nature

www.nature.com/nature

Vol 454 | Issue no. 7203 | 24 July 2008

China's challenges

By almost every measure, China's growth is extraordinary. But behind the astonishing statistics is a more complex reality.

CHINA

Discussions of China's emergence as a superpower often focus on matters of scale. This is understandable. China's borders encompass more than 1.3 billion people — one in every five humans on the planet — and stunningly diverse terrain, from the Yellow River plain in the east to the Himalayan plateau in the west. In science and technology, China now generates more publications than any other country bar the United States, and ranks third in the number of doctoral degrees it awards. One can take almost any measure and find an extreme in China. Where else would authorities even consider a plan to redistribute water resources by diverting major rivers for more than a thousand kilometres?

But these gargantuan figures may not mean quite what they seem to. China accounts for a smaller proportion of the world's population than it did in the seventeenth century. Many analysts concur that the nation's current economic growth is in a boom phase that cannot last. Most importantly, as the articles in this special issue illustrate, the image of China as a monolithic juggernaut hides a more complex and interesting reality. On page 412, Rogers Hollingsworth and his colleagues argue that the burgeoning of Chinese science does not necessarily mean it will replace the United States as the new hegemon, but rather that it will find a prominent role among a more diverse global research community in which no nation dominates.

"China's current success story continues to be characterized by a canny pragmatism."

Moreover, it is not clear whether China's growing strength in science — which increasingly belies the lazy notion that research in Asian countries lacks originality — will automatically make Chinese institutions major players at all the established frontiers, from drug development to nanotechnology or space science. Nations have different priorities, and this is especially true for those whose economic and technological development is relatively recent. On page 398, Lan Xue outlines the hazards of simply competing on the basis of an agenda determined by previous scientific superpowers, with its unspoken rules about which are the most important areas of research and where results should be published. If China were to decide that its interests lie in, say, massive investment in clean energy production — a matter of national urgency, for which it might be unwise to rely on the leadership of under-funded work in the West — it could both address its own needs and command immense respect on a global stage.

Indeed, the global problems that would be tackled by a Chinese focus on domestic priorities make for an almost perfect wish-list. These priorities include water conservation and water pollution treatment, earthquake forecasting and earthquake-resistant building technologies, flood management and drought-resistant crops — all of which would find widespread application elsewhere. Likewise with health: China's domestic challenges are also the world's challenges. Already, for example, the antimalarial drug artemisinin is one of the most celebrated bounties of Chinese herbal medicine. Avian flu threatens to be China's home-grown epidemic. And the spread of AIDS is now acknowledged as a national issue, especially after the scandal of HIV infections of peasant blood donors in Henan province in the 1990s undermined official denial. When Premier Wen Jiabao was photographed in 2003 shaking hands with an AIDS patient, it seemed clear that the government was at last facing up to the problem.

One of the biggest questions for outside observers is to what extent the social, cultural and political milieu is shaping, and will continue to shape, the very practice of doing science in China. That legislation was proposed last year to make it acceptable for researchers to admit their failures (see *Nature* **449,** 12; 2007) suggests that there are deeply ingrained

For podcast and more online extras see www.nature.com/news/specials/china/

IMAGE SOURCE/GETTY IMAGES

misconceptions at an institutional level about how science works — misconceptions that stifle risk-taking and promote narrow conformity. But some suspect that such cosmetic efforts do little to address the problems created by a strongly hierarchical research culture, where an immense pressure to succeed might be seen as a precondition for the sorts of abuses evident in the case of disgraced cloning researcher Woo Suk Hwang in South Korea.

An even deeper question is whether a truly vibrant scientific culture is possible without a more widespread societal commitment to free expression. The right to challenge authority, and to doubt everything, is central to scientific enquiry. And no country can be a major scientific player in the modern world unless its scientists can collaborate with researchers from elsewhere. A poor record on human rights will not make this impossible — but it will make it more difficult. Scientists do, largely, have a commitment to human rights, and will be happier working with colleagues who share that commitment.

Perhaps it is good news that China's current success story continues to be characterized by a canny pragmatism. Granted, this attitude can sometimes make it seem as though everything is motivated by the economic bottom line. For example, although it would be too cynical to suggest that global warming and environmental degradation are now being taken seriously only because they eat into China's gross domestic product, that is surely one big reason for the concern. Yet, motivated by that same pragmatism, the Chinese authorities are increasingly recognizing that getting the best value from its scientists means providing them with adequate funds and minimal interference, even if this sees them straying 'off-message'. Many outside scientists have been surprised to find that Chinese graduate students and postdocs are now quite willing to challenge their professors. Exaggerated deference to authority is clearly on the wane in China's younger generation of scientists — and who knows how far that pragmatic liberalization will go?

In the meantime, the rest of the world can surely benefit from the self-confidence that will make China a source not just of skilled, hard-working postdocs, but of a new agenda, informed by a tradition of innovation of almost unparalleled antiquity and sophistication. ■

Mind the gaps

The incoming US administration can and should reverse the neglect of Earth observations.

At many places around the world, it is possible to feel the climate changing: the ice cracking, the soil waking earlier in the spring. Perhaps such feelings are merely rooted in a heightened awareness of global warming. Whatever the case, the true significance of such localized experiences can be assessed only through extensive data on key Earth processes over time.

Given the effects of those processes, it would be smart and responsible for policy-makers and scientists to focus on strengthening the international strategy for collecting key Earth observations for the foreseeable future. After all, data on such variables as soil moisture, wind speed and direction at various altitudes, and atmospheric temperature and pressure are essential for improving climate models in the future, and for informing mitigation efforts. Moreover, many of the programmatic elements for global data-gathering are already in place — one notable example being the consortium pursuing the Global Earth Observation System of Systems (GEOSS), which comprises 74 nations plus the European Commission.

And yet the United States, potentially one of the central participants in any such international system, has allowed short-term budgetary considerations to direct too many of its decisions. In June 2006, for example, cost over-runs caught up with the National Polar-orbiting Operational Environmental Satellite System, a proposed joint military and civil satellite programme intended to replace all US weather satellites. Several long-term climate data sensors were axed or pared down, including those measuring the variables mentioned above. A few months later, the Geostationary Operational Environmental Satellites programme got the same treatment.

Earlier this month, the National Academies' National Research Council (NRC) released a report that suggests ways to save from oblivion the data those sensors would have collected. The council identifies important sensors to be restored to the missions from which they were cut, and recommends that others should be put on new satellites or should hitch rides on other scheduled missions.

After presenting this patchwork solution, the report stresses the importance of a long-term strategy for climate observation. This section should be required reading for anyone hoping for a political appointment at NASA, the National Oceanic and Atmospheric Administration (NOAA) or the US Geological Survey in 2009. In fact, the hopefuls should add the NRC's decadal survey *Earth Science and Applications from Space* to their reading list. What is clear from both documents is that the United States does not have a unified strategy for collecting these observations, and that the three agencies involved have not been able to avoid gaps in data or unnecessary duplication of data gathered by other nations.

The White House will need to exert pressure to make such a strategy a high priority. Neither of the two presidential candidates, John McCain and Barack Obama, has said much, if anything, about Earth observations. Providing for continuous high-quality climate data would be a substantial legacy, serving the interests of both US citizens and the rest of the world for decades to come.

The NRC's suggestions should be taken up immediately, and the next US president should move quickly to appoint directors to NOAA and NASA who see Earth monitoring as a priority. Commendably, the current head of NOAA, Conrad Lautenbacher, has been a driving force behind GEOSS. The occasionally sceptical remarks about climate change from NASA administrator Michael Griffin make him less of an example to follow.

At the very least, the new agency heads should commit to the monitoring of the essential climate variables outlined by the Global Climate Observing System. They should work with other countries through GEOSS to make sure that data are not needlessly duplicated. Satellites can provoke secrecy and competition between nations that, instead, must pull together to monitor the well-being of the planet on which they all depend. ■

misconceptions at an institutional level about how science works — misconceptions that stifle risk-taking and promote narrow conformity. But some suspect that such cosmetic efforts do little to address the problems created by a strongly hierarchical research culture, where an immense pressure to succeed might be seen as a precondition for the sorts of abuses evident in the case of disgraced cloning researcher Woo Suk Hwang in South Korea.

An even deeper question is whether a truly vibrant scientific culture is possible without a more widespread societal commitment to free expression. The right to challenge authority, and to doubt everything, is central to scientific enquiry. And no country can be a major scientific player in the modern world unless its scientists can collaborate with researchers from elsewhere. A poor record on human rights will not make this impossible — but it will make it more difficult. Scientists do, largely, have a commitment to human rights, and will be happier working with colleagues who share that commitment.

Perhaps it is good news that China's current success story continues to be characterized by a canny pragmatism. Granted, this attitude can sometimes make it seem as though everything is motivated by the economic bottom line. For example, although it would be too cynical to suggest that global warming and environmental degradation are now being taken seriously only because they eat into China's gross domestic product, that is surely one big reason for the concern. Yet, motivated by that same pragmatism, the Chinese authorities are increasingly recognizing that getting the best value from its scientists means providing them with adequate funds and minimal interference, even if this sees them straying 'off-message'. Many outside scientists have been surprised to find that Chinese graduate students and postdocs are now quite willing to challenge their professors. Exaggerated deference to authority is clearly on the wane in China's younger generation of scientists — and who knows how far that pragmatic liberalization will go?

In the meantime, the rest of the world can surely benefit from the self-confidence that will make China a source not just of skilled, hard-working postdocs, but of a new agenda, informed by a tradition of innovation of almost unparalleled antiquity and sophistication. ■

Mind the gaps

The incoming US administration can and should reverse the neglect of Earth observations.

At many places around the world, it is possible to feel the climate changing: the ice cracking, the soil waking earlier in the spring. Perhaps such feelings are merely rooted in a heightened awareness of global warming. Whatever the case, the true significance of such localized experiences can be assessed only through extensive data on key Earth processes over time.

Given the effects of those processes, it would be smart and responsible for policy-makers and scientists to focus on strengthening the international strategy for collecting key Earth observations for the foreseeable future. After all, data on such variables as soil moisture, wind speed and direction at various altitudes, and atmospheric temperature and pressure are essential for improving climate models in the future, and for informing mitigation efforts. Moreover, many of the programmatic elements for global data-gathering are already in place — one notable example being the consortium pursuing the Global Earth Observation System of Systems (GEOSS), which comprises 74 nations plus the European Commission.

And yet the United States, potentially one of the central participants in any such international system, has allowed short-term budgetary considerations to direct too many of its decisions. In June 2006, for example, cost over-runs caught up with the National Polar-orbiting Operational Environmental Satellite System, a proposed joint military and civil satellite programme intended to replace all US weather satellites. Several long-term climate data sensors were axed or pared down, including those measuring the variables mentioned above. A few months later, the Geostationary Operational Environmental Satellites programme got the same treatment.

Earlier this month, the National Academies' National Research Council (NRC) released a report that suggests ways to save from oblivion the data those sensors would have collected. The council identifies important sensors to be restored to the missions from which they were cut, and recommends that others should be put on new satellites or should hitch rides on other scheduled missions.

After presenting this patchwork solution, the report stresses the importance of a long-term strategy for climate observation. This section should be required reading for anyone hoping for a political appointment at NASA, the National Oceanic and Atmospheric Administration (NOAA) or the US Geological Survey in 2009. In fact, the hopefuls should add the NRC's decadal survey *Earth Science and Applications from Space* to their reading list. What is clear from both documents is that the United States does not have a unified strategy for collecting these observations, and that the three agencies involved have not been able to avoid gaps in data or unnecessary duplication of data gathered by other nations.

The White House will need to exert pressure to make such a strategy a high priority. Neither of the two presidential candidates, John McCain and Barack Obama, has said much, if anything, about Earth observations. Providing for continuous high-quality climate data would be a substantial legacy, serving the interests of both US citizens and the rest of the world for decades to come.

The NRC's suggestions should be taken up immediately, and the next US president should move quickly to appoint directors to NOAA and NASA who see Earth monitoring as a priority. Commendably, the current head of NOAA, Conrad Lautenbacher, has been a driving force behind GEOSS. The occasionally sceptical remarks about climate change from NASA administrator Michael Griffin make him less of an example to follow.

At the very least, the new agency heads should commit to the monitoring of the essential climate variables outlined by the Global Climate Observing System. They should work with other countries through GEOSS to make sure that data are not needlessly duplicated. Satellites can provoke secrecy and competition between nations that, instead, must pull together to monitor the well-being of the planet on which they all depend. ■

RESEARCH HIGHLIGHTS

Bird's-nose view

Proc. R. Soc. B doi:10.1098/rspb.2008.0607 (2008)
Smell may be much more important to the way birds perceive their surroundings than biologists have thought. A study of nine species of bird from seven orders found, in all cases, that the majority of olfactory-receptor genes were probably functional, report Silke Steiger of the Max Planck Institute for Ornithology in Starnberg, Germany, and her co-workers. The only previous estimate — from a draft genomic sequence of the red jungle fowl (*Gallus gallus*) — put that proportion at just 15%.

The total number of working olfactory-receptor genes that an animal has probably indicates how many different scents it can distinguish. Of the species in this sample, the kakapo (*Strigops habroptilus*, pictured), which forages at night, had the most 'smell' genes, 82% of which probably contribute to this bird's sense of smell.

C. COURTEAU/NATUREPL.COM

PHYSICS

Parting a cloud

Appl. Phys. Lett. **92,** 254102 (2008)
A team of researchers has made three-dimensional 'atom chips' that give unprecedented control over Bose–Einstein condensates (BECs) — clouds of extremely cold atoms that all share the same quantum state.

Thorsten Schumm at Vienna University of Technology and his colleagues used ultraviolet light and electron beams to pattern multiple wiring layers, separated by insulators, onto a semiconductor. By running currents through the wires, the team created magnetic potentials able to hold and manipulate BECs.

For instance, they can split a BEC in two and perform experiments on its halves. They believe that the work might lead to highly sensitive magnetometers and applications in quantum information technology.

B. FASANI/CORBIS

ACOUSTICS

Chuckle vision

J. Acoust. Soc. Am. **124,** 472–483 (2008)
Laughter is considered to be a reflex action, an adaptive tension-reliever with analogues in many non-human species. That congenitally deaf people laugh out loud supports this theory. But do they produce the same sounds as those who hear normally?

Maja Makagon of Cornell University in Ithaca, New York, and her colleagues showed congenitally deaf volunteers clips of films such as *Mr Bean* and *The Naked Gun*, and compared the acoustic properties of their laughter with that of unimpaired controls. The quality of the sound was remarkably similar; the differences in the sound-waves' shapes were more consistent with deaf people having less vocal-muscle control than with hearers having learned how laughing 'should sound'.

The deaf volunteers laughed more quietly, perhaps owing to social conditioning that led them to lower vocal volume overall.

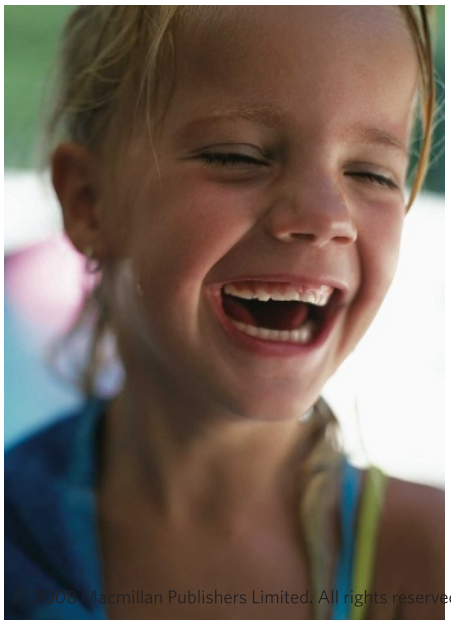
PLANT SCIENCES

Poisonous grains

Proc. Natl Acad. Sci. USA doi:10.1073/pnas.0802361105 (2008)
Rice is efficient, indeed disconcertingly so, at assimilating arsenic from the soils of paddy fields. But how it does this has been unclear. Now Fang-Jie Zhao at Rothamsted Research in Harpenden, UK, Jian Feng Ma at Okayama University in Japan and their colleagues have discovered that it is taken into the plant as though it were silicon.

They found that two transporter proteins belonging to the family known as aquaporins enable arsenite to move from rice's soggy surroundings into its vascular system. Mutations in the genes encoding either of these proteins reduced arsenite uptake by the roots and the amount of arsenic that accumulated in shoots and grains.

The authors hope that different versions of these genes exist that favour silicon transport over that of arsenite. If so, rice carrying such versions could be planted in regions of the world where arsenic poisoning is a problem.



PHYSICS

Gravity up close

Phys. Rev. D **78,** 022002 (2008)
Gravity is the weakest and least well understood of the four fundamental forces. It behaves well over large distances. But many theorists suspect that undiscovered particles or extra dimensions might cause its observed behaviour to break down over very short distances — which might help to reconcile gravity with the three other forces.

Current approaches will hold a little while longer, however, thanks to Andrew Geraci and his colleagues at Stanford University in California, who have made the most precise measurements yet of gravity over 10 micrometres. They found no anomalies.

The researchers placed a 1.5-microgram gold cuboid on a silicon cantilever a quarter of a millimetre long, rather like a diver on a

RESEARCH HIGHLIGHTS

Bird's-nose view

Proc. R. Soc. B doi:10.1098/rspb.2008.0607 (2008)
Smell may be much more important to the way birds perceive their surroundings than biologists have thought. A study of nine species of bird from seven orders found, in all cases, that the majority of olfactory-receptor genes were probably functional, report Silke Steiger of the Max Planck Institute for Ornithology in Starnberg, Germany, and her co-workers. The only previous estimate — from a draft genomic sequence of the red jungle fowl (*Gallus gallus*) — put that proportion at just 15%.

The total number of working olfactory-receptor genes that an animal has probably indicates how many different scents it can distinguish. Of the species in this sample, the kakapo (*Strigops habroptilus*, pictured), which forages at night, had the most 'smell' genes, 82% of which probably contribute to this bird's sense of smell.

C. COURTEAU/NATUREPL.COM

PHYSICS

Parting a cloud

Appl. Phys. Lett. **92,** 254102 (2008)
A team of researchers has made three-dimensional 'atom chips' that give unprecedented control over Bose–Einstein condensates (BECs) — clouds of extremely cold atoms that all share the same quantum state.

Thorsten Schumm at Vienna University of Technology and his colleagues used ultraviolet light and electron beams to pattern multiple wiring layers, separated by insulators, onto a semiconductor. By running currents through the wires, the team created magnetic potentials able to hold and manipulate BECs.

For instance, they can split a BEC in two and perform experiments on its halves. They believe that the work might lead to highly sensitive magnetometers and applications in quantum information technology.

B. FASANI/CORBIS

ACOUSTICS

Chuckle vision

J. Acoust. Soc. Am. **124,** 472–483 (2008)
Laughter is considered to be a reflex action, an adaptive tension-reliever with analogues in many non-human species. That congenitally deaf people laugh out loud supports this theory. But do they produce the same sounds as those who hear normally?

Maja Makagon of Cornell University in Ithaca, New York, and her colleagues showed congenitally deaf volunteers clips of films such as *Mr Bean* and *The Naked Gun*, and compared the acoustic properties of their laughter with that of unimpaired controls. The quality of the sound was remarkably similar; the differences in the sound-waves' shapes were more consistent with deaf people having less vocal-muscle control than with hearers having learned how laughing 'should sound'.

The deaf volunteers laughed more quietly, perhaps owing to social conditioning that led them to lower vocal volume overall.

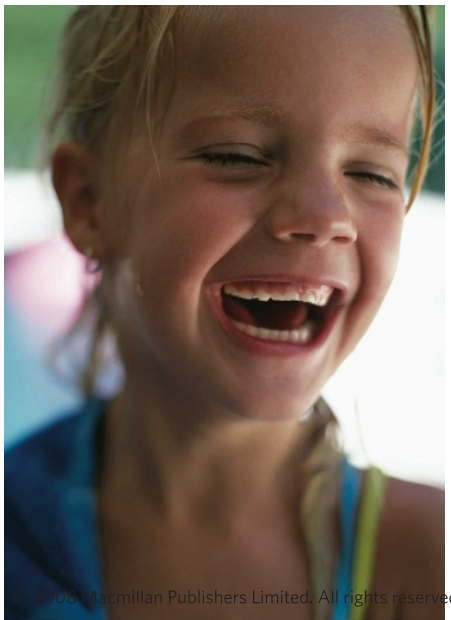
PLANT SCIENCES

Poisonous grains

Proc. Natl Acad. Sci. USA doi:10.1073/pnas.0802361105 (2008)
Rice is efficient, indeed disconcertingly so, at assimilating arsenic from the soils of paddy fields. But how it does this has been unclear. Now Fang-Jie Zhao at Rothamsted Research in Harpenden, UK, Jian Feng Ma at Okayama University in Japan and their colleagues have discovered that it is taken into the plant as though it were silicon.

They found that two transporter proteins belonging to the family known as aquaporins enable arsenite to move from rice's soggy surroundings into its vascular system. Mutations in the genes encoding either of these proteins reduced arsenite uptake by the roots and the amount of arsenic that accumulated in shoots and grains.

The authors hope that different versions of these genes exist that favour silicon transport over that of arsenite. If so, rice carrying such versions could be planted in regions of the world where arsenic poisoning is a problem.



PHYSICS

Gravity up close

Phys. Rev. D **78,** 022002 (2008)
Gravity is the weakest and least well understood of the four fundamental forces. It behaves well over large distances. But many theorists suspect that undiscovered particles or extra dimensions might cause its observed behaviour to break down over very short distances — which might help to reconcile gravity with the three other forces.

Current approaches will hold a little while longer, however, thanks to Andrew Geraci and his colleagues at Stanford University in California, who have made the most precise measurements yet of gravity over 10 micrometres. They found no anomalies.

The researchers placed a 1.5-microgram gold cuboid on a silicon cantilever a quarter of a millimetre long, rather like a diver on a

diving-board. They then measured the extra bend, due to gravity, when a second mass was temporarily brought directly beneath the gold.

NEUROSCIENCE

Location, location, location

Neuron **59,** 125–137 (2008)

Researchers studying anaesthetized adult gerbils fitted with earphones report that the neurotransmitter GABA calibrates the processing system that locates a sound's origin.

Ursula Koch and Anna Magnusson of LMU Munich in Germany and their co-workers considered the lateral superior olive (LSO), a nucleus in the gerbil brainstem where information from both ears converges. They played different sound volumes through the right and left earphones and administered chemicals that stimulate or block GABA receptors. This revealed that GABA released by neurons in the LSO adjusts the balance of excitation and inhibition experienced by the same neurons as a result of signals from each ear.

Excitatory nerve terminals seemed to be more strongly affected by GABA, which suggests that neurons in the LSO tend to 'turn down' excitatory input. This would increase auditory sensitivity on the side of the animal that a sound is coming from.

ASTRONOMY

Bright origins

Astrophys. J. **681,** 1035–1045 (2008)

Astronomers have found that vast stores of hot gas in the areas between clusters of gravitationally bound galaxies do form stars, though not many. The gas falls into one of the cluster's bright central galaxies, where it cools and condenses enough for star formation. This process was thought to be negligible in the present-day Universe.

Christopher O'Dea from the Rochester Institute of Technology in New York and his colleagues considered data from 62 of these central galaxies, from which they estimate that 1–10% of the gas contributes to star birth. X-ray emissions served as a proxy for the amount of hot gas falling in, and infrared emissions as a proxy for new stars being formed. Some mechanism, the authors suggest, keeps the gas from cooling completely — perhaps a supermassive black hole in the galactic core, or the new stars themselves.

INFECTIOUS DISEASE

DARC matters

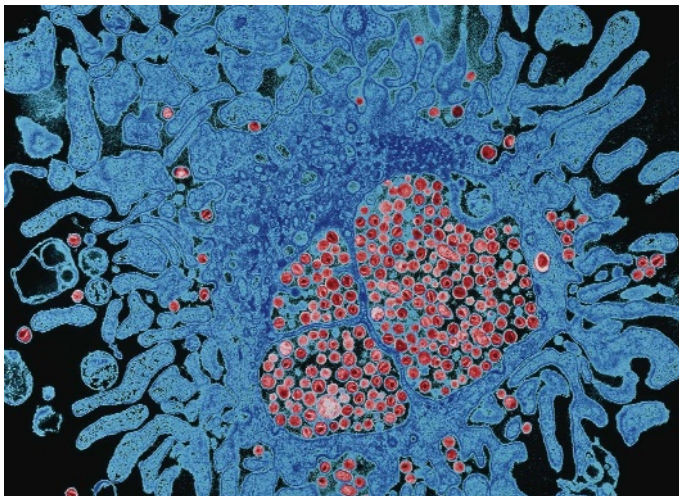
Cell Host Microbe **4,** 52–62 (2008)

A mutation that makes Africans resistant to a form of malaria renders them more vulnerable to HIV infection, researchers have found.

The mutation halts the expression of the protein DARC in red blood cells, where it normally occurs on the surface. Almost all black Africans carry this mutation, which confers resistance to the benign, recurring malaria caused by the parasite *Plasmodium vivax*.

Sunil Ahuja of the University of Texas Health Science Center in San Antonio and his colleagues analysed blood samples from more than 3,400 African Americans and discovered that the DARC mutation is associated with a 40% increase in the risk of acquiring HIV. However, HIV-infected participants with the DARC mutation also survived an average of two years longer than those without it.

The image (below) shows an immune cell known as a T lymphocyte full of newly manufactured HIV particles (red).



H. GELDERBLOM/GETTY IMAGES

CHEMISTRY

Easy bonding

Angew. Chem. Int. Edn doi:10.1002/anie.200802164 (2008)

Many drugs contain compounds with fluorine–carbon bonds, as do tracers used in positron-emission tomography (PET), a medical imaging technique. Producing these compounds is tricky and involves harsh conditions. Now, Tobias Ritter and his colleagues at Harvard University in Cambridge, Massachusetts, have worked out how to perform the fluorination reaction at room temperature.

They developed a palladium catalyst that can replace a boronic acid group on an aromatic ring with fluorine. The catalyst has nitrogen-containing ligands that make it resistant to attack from aggressive fluorination reagents. Other chemical groups on the ring do not interfere with the reaction, and the carbon–fluorine bond forms in the final step. That is important for making PET tracers because the fluorine isotopes used for PET have short half-lives.

MOLECULAR BIOLOGY

WHAMM!

Cell **134,** 148–161 (2008)

A protein called WHAMM helps shuttle other proteins between compartments in mammalian cells by interacting with two components of the cell skeleton, researchers have found.

Matthew Welch, Kenneth Campellone and their colleagues at the University of California, Berkeley, found that WHAMM mediates the transport of proteins between the endoplasmic reticulum and the Golgi apparatus. The researchers mapped distinct regions of the protein that interact with the membranes of the Golgi, and with two constituents of the cell's internal skeleton: actin and microtubules.

Both raising and lowering the amount of WHAMM in human cells disrupted the Golgi's structure and interfered with the transport of a viral protein from the endoplasmic reticulum to the Golgi.

GENETICS

DNA potholes

Proc. Natl Acad. Sci. USA **105,** 9936–9941 (2008)

In living cells, palindromes in a DNA sequence often stall the DNA replication machinery when their two halves bind, making the strand loop outwards.

Such arrangements, in which similar or identical sequences sit close to each other but run in opposite directions, are hotspots for chromosome breaks that can cause disease. Using gel electrophoresis to analyse DNA at various stages of its copying, Sergei Mirkin of Tufts University in Medford, Massachusetts, and his colleagues showed that hairpin structures are made this way in living bacterial, yeast and primate cells.

The researchers think that when a hairpin forms, the lagging strand is left uncopied. This makes it more prone to breakage, and thus at greater risk of elimination from the genome.

diving-board. They then measured the extra bend, due to gravity, when a second mass was temporarily brought directly beneath the gold.

NEUROSCIENCE

Location, location, location

Neuron **59,** 125–137 (2008)

Researchers studying anaesthetized adult gerbils fitted with earphones report that the neurotransmitter GABA calibrates the processing system that locates a sound's origin.

Ursula Koch and Anna Magnusson of LMU Munich in Germany and their co-workers considered the lateral superior olive (LSO), a nucleus in the gerbil brainstem where information from both ears converges. They played different sound volumes through the right and left earphones and administered chemicals that stimulate or block GABA receptors. This revealed that GABA released by neurons in the LSO adjusts the balance of excitation and inhibition experienced by the same neurons as a result of signals from each ear.

Excitatory nerve terminals seemed to be more strongly affected by GABA, which suggests that neurons in the LSO tend to 'turn down' excitatory input. This would increase auditory sensitivity on the side of the animal that a sound is coming from.

ASTRONOMY

Bright origins

Astrophys. J. **681,** 1035–1045 (2008)

Astronomers have found that vast stores of hot gas in the areas between clusters of gravitationally bound galaxies do form stars, though not many. The gas falls into one of the cluster's bright central galaxies, where it cools and condenses enough for star formation. This process was thought to be negligible in the present-day Universe.

Christopher O'Dea from the Rochester Institute of Technology in New York and his colleagues considered data from 62 of these central galaxies, from which they estimate that 1–10% of the gas contributes to star birth. X-ray emissions served as a proxy for the amount of hot gas falling in, and infrared emissions as a proxy for new stars being formed. Some mechanism, the authors suggest, keeps the gas from cooling completely — perhaps a supermassive black hole in the galactic core, or the new stars themselves.

INFECTIOUS DISEASE

DARC matters

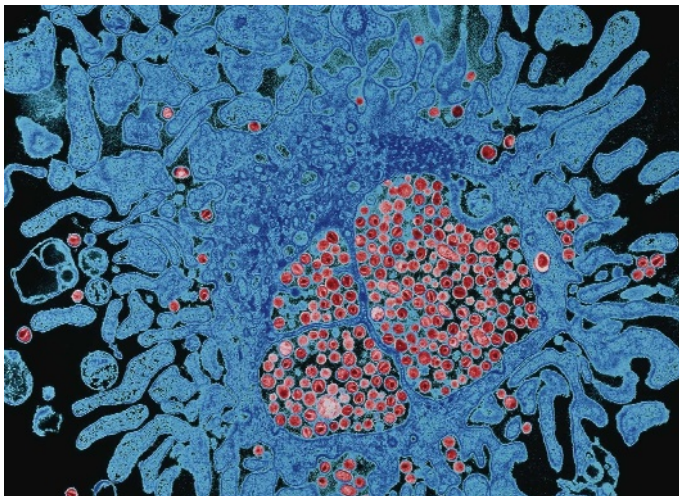
Cell Host Microbe **4,** 52–62 (2008)

A mutation that makes Africans resistant to a form of malaria renders them more vulnerable to HIV infection, researchers have found.

The mutation halts the expression of the protein DARC in red blood cells, where it normally occurs on the surface. Almost all black Africans carry this mutation, which confers resistance to the benign, recurring malaria caused by the parasite *Plasmodium vivax*.

Sunil Ahuja of the University of Texas Health Science Center in San Antonio and his colleagues analysed blood samples from more than 3,400 African Americans and discovered that the DARC mutation is associated with a 40% increase in the risk of acquiring HIV. However, HIV-infected participants with the DARC mutation also survived an average of two years longer than those without it.

The image (below) shows an immune cell known as a T lymphocyte full of newly manufactured HIV particles (red).



H. GELDERBLOM/GETTY IMAGES

CHEMISTRY

Easy bonding

Angew. Chem. Int. Edn doi:10.1002/anie.200802164 (2008)

Many drugs contain compounds with fluorine–carbon bonds, as do tracers used in positron-emission tomography (PET), a medical imaging technique. Producing these compounds is tricky and involves harsh conditions. Now, Tobias Ritter and his colleagues at Harvard University in Cambridge, Massachusetts, have worked out how to perform the fluorination reaction at room temperature.

They developed a palladium catalyst that can replace a boronic acid group on an aromatic ring with fluorine. The catalyst has nitrogen-containing ligands that make it resistant to attack from aggressive fluorination reagents. Other chemical groups on the ring do not interfere with the reaction, and the carbon–fluorine bond forms in the final step. That is important for making PET tracers because the fluorine isotopes used for PET have short half-lives.

MOLECULAR BIOLOGY

WHAMM!

Cell **134,** 148–161 (2008)

A protein called WHAMM helps shuttle other proteins between compartments in mammalian cells by interacting with two components of the cell skeleton, researchers have found.

Matthew Welch, Kenneth Campellone and their colleagues at the University of California, Berkeley, found that WHAMM mediates the transport of proteins between the endoplasmic reticulum and the Golgi apparatus. The researchers mapped distinct regions of the protein that interact with the membranes of the Golgi, and with two constituents of the cell's internal skeleton: actin and microtubules.

Both raising and lowering the amount of WHAMM in human cells disrupted the Golgi's structure and interfered with the transport of a viral protein from the endoplasmic reticulum to the Golgi.

GENETICS

DNA potholes

Proc. Natl Acad. Sci. USA **105,** 9936–9941 (2008)

In living cells, palindromes in a DNA sequence often stall the DNA replication machinery when their two halves bind, making the strand loop outwards.

Such arrangements, in which similar or identical sequences sit close to each other but run in opposite directions, are hotspots for chromosome breaks that can cause disease. Using gel electrophoresis to analyse DNA at various stages of its copying, Sergei Mirkin of Tufts University in Medford, Massachusetts, and his colleagues showed that hairpin structures are made this way in living bacterial, yeast and primate cells.

The researchers think that when a hairpin forms, the lagging strand is left uncopied. This makes it more prone to breakage, and thus at greater risk of elimination from the genome.

NEWS

Oil cost hits ship studies

J. ZIEGLER

The *Polarstern* research vessel in Antarctica. Its fuel costs have increased fivefold since 2003.

Many of the research projects launched as part of the International Polar Year (IPY), which runs from March 2007 to March 2009, are under threat because of the steep rise in marine-fuel costs. Hundreds of Arctic and Antarctic scientists face uncertainty as polar science programmes worldwide are curtailed, postponed or cancelled.

The price of a barrel of oil has more than doubled since March 2007, from US$60 to $140 now. High energy costs are a problem for research in most fields, but logistically complicated research operations in remote polar regions are more affected than, say, big physics experiments.

"We have reached a point where the collapse of some of our activities is looming on the horizon," says Karin Lochte, director of the Alfred Wegener Institute for Polar and Marine Research (AWI) in Bremerhaven, Germany, which operates the research icebreaker *Polarstern*, Europe's largest scientific vessel.

Icebreakers are usually fuelled by marine diesel oil (MDO), a cleaner and more expensive fuel than the heavy oil used by normal cargo ships. The average price for MDO has increased fivefold since 2003, from $250 to $1,300 per metric tonne (equivalent to around 1,200 litres of diesel). Since January, the price has increased by almost $550 per tonne (see graph).

Operating the *Polarstern*, which is usually at sea for around 320 days per year, currently costs around $100,000 per day, with fuel now accounting for half of that cost. Logistics experts at the AWI estimate that by the end of the year the institute will have exceeded, by more than $5 million, its $30-million budget for operating the ship. The German science ministry, from which the AWI receives 90% of its budget, has rejected requests to cover the extra costs.

"If fuel prices remain as high, we will be forced to cut projects and substantially reduce days at sea," says Lochte. One option, she says, is to cancel the *Polarstern*'s entire Arctic programme next year. The ship could also be temporarily chartered to non-scientific users, she says.

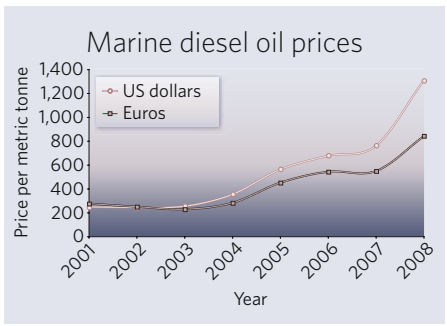
More than 100 *Polarstern* scientists could be hit by cutbacks. Arctic projects at risk include a wide variety of geophysical, oceanographic and biological research, such as sample-taking in the AWI's Hausgarten, a deep-sea long-term observatory in the eastern Fram Strait, between Greenland and Spitsbergen, and measurements of water flux through the strait.

Rising fuel costs threaten researchers from all countries involved in polar research. David Barber, chief scientist of the $40-million Canadian-led Circumpolar Flaw Lead study — the biggest single IPY project — fears that such expensive expeditions might soon become unaffordable.

In the United States, increased fuel and transportation costs have caused a $30-million shortfall in this fiscal year's $350-million Antarctica budget of the National Science Foundation's Office of Polar Programs (OPP). The agency has asked the government for a budget increase to compensate for the extra costs. But Congress is unlikely to pass the request before the next US administration is in office.

"We have to figure out how to get through," says Karl Erb, director of the OPP. "We will make every effort to keep the science going and not cancel projects, but I do know we have to reduce or defer some of the things we had planned next year."

The OPP is operating four icebreakers and a fleet of cargo aircraft in support of Antarctic research activities. Although ship operations will be maintained, the number of transport flights from New Zealand to Antarctica will be reduced, Erb says. A geophysical study of the Gamburtsev Mountains in eastern Antarctica — a large international IPY collaboration — is one of the projects that might get less air support, he says.



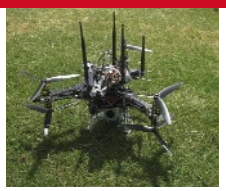
And the deployment of seismic sensors for the Polar Earth Observing Network (POLENET) project, an IPY collaboration with aims such as measuring the uplift of Earth's crust as a result of the West Antarctic ice sheet losing mass, may have to wait another year, says Erb.

Severe concerns about soaring costs were also raised last week in Moscow during the biennial meeting of the Scientific Committee on Antarctic Research (SCAR). "The common tenor is that painful cuts are unavoidable," says Lochte, the German delegate to SCAR. To mitigate the fallout, the group is now discussing saving money by improving international cooperation. Joint logistics would help to reduce the number of aircraft flights and ship operations required for supplying and maintaining Antarctic stations.

Reduced ship time could also be partly compensated for by a more widespread use of buoys, autonomous underwater vehicles and unmanned aircraft, says Lochte.

Alternatives to diesel-fuelled ship engines, perhaps on the basis of hydrogen-powered fuel cells, are in an experimental stage of development at best. Nuclear-driven research ships, such as the fleet of nuclear icebreakers that Russia has in use, are also an option — but one that most governments consider environmentally and politically too risky. ■

Quirin Schiermeier



S. PRIOR/MIDDLESEX UNIV.

FLYING ROBOTS HAVE THEIR WINGS CLIPPED
Science misses out on unmanned aerial vehicles thanks to poor regulation.
www.nature.com/news

Spinal cord revealed in free gene map

The Allen Institute for Brain Science has released the first data from its ambitious project to map the spinal cord. When completed early next year, the freely accessible atlas will chart the expression patterns of at least 18,500 genes throughout the spinal cord of juvenile and adult mice. The first data, released on 17 July, cover more than 2,000 of those genes.

The atlas is a follow-up to the Allen Brain Atlas, a virtual map of the expression of 20,000 genes in the mouse brain, which was completed in 2006 (E. S. Lein *et al. Nature* **445,** 168–176; 2007). Both projects come courtesy of Paul Allen, co-founder of Microsoft, who has a long-held fascination with brain circuitry.

"The resolution is really quite phenomenal."

The latest genetic atlas is being constructed from 20-micrometre-thick sections of spinal cord, taken at millimetre intervals. Researchers will be able to compare gene activity in a four-day-old mouse with that in an adult, and to zoom in on pictures resolvable down to one micrometre per pixel of screen — individual nerve cells are upwards of 10 micrometres in diameter.

"The Allen Brain Atlas, and now the Allen Spinal Cord Atlas, look comprehensively at what's going on at the cellular level to a degree no one has done before," says Allan Jones, chief scientific officer of the institute, which is based in Seattle, Washington. The $2.3-million spinal-cord project could yield clues about, for example, the genes involved in neural regeneration — $600,000 of the costs came from a coalition that includes the Paralyzed Veterans of America, the US Amyotrophic Lateral Sclerosis (ALS) Association and the US National Multiple Sclerosis Society.

"Motor neurons appear enormous when I zoom in," says neuroscientist Jane Roskams of the University of British Columbia in Vancouver, Canada, who has tried out the atlas. "The images are extremely clear. You can see every single cell, and you can tell the relative positions of cells in different layers," she says. "The resolution is really quite phenomenal, giving the detail you need to assess whether a gene is contributing to the function of a particular cell type."

"It's like looking down on the surface of the Earth and suddenly having the distribution of all the mineral resources beneath the surface revealed to you without having do any digging," says David Anderson of the California Institute of Technology in Pasadena, who is a scientific adviser to the Allen Institute.

Steve McMahon of King's College London, who studies chronic pain mediated by the spinal cord, agrees that the new atlas will be "fantastically useful". But he cautions: "It's an atlas of the normal. It won't tell us about motor neuron disease or multiple sclerosis or chronic pain." In a large number of spinal-cord pathologies, there is a dramatic change in both cell composition and gene expression in the diseased or damaged tissue.

Roskams counters that understanding the spinal cord's healthy state is a necessary first step in trying to fix diseased or injured cells. "If you are going to have a genetic model of ALS that has dying motor neurons and compromised glial cells, and you want to try to restore that model to a healthy state, you've got to have a pretty good idea of what you're restoring it to."

But the most clinically useful information may have to wait for the human project: the institute hopes to complete the Allen Human Brain Atlas by 2012. ■

Meredith Wadman

Think tank reveals plan to manage tropical forests

A high-profile group of thinkers has come up with a straightforward way to integrate long-term forest management into an international agreement on halting deforestation.

It is not clear whether the proposal — released on 18 July by the Terrestrial Carbon Group of international scientists, economists and land-policy experts — has political legs.

The idea is to 'lump together' all of the carbon locked up in tropical forests and then allow all countries to cash in on forest protection by trading carbon credits, regardless of whether logging is currently a problem within their borders. The proposal differs from the leading framework under the current United Nations (UN) proposal, which would establish baseline deforestation rates for each country and then allow them to sell carbon credits into an international market if they can reduce the rate of deforestation.

Countries that do not currently have problems with deforestation stand to gain nothing under the UN system, which would mainly benefit countries such as Brazil and Indonesia. Many fear that it could leave the rest of the tropics exposed to logging pressure in the future.

Dan Nepstad, a deforestation expert who is a member of the Terrestrial Carbon Group, says that the group's proposal emphasizes the need for a comprehensive solution for dealing with all tropical forests. "My hope is that it sends a signal to India, Costa Rica, China and other places that there will ultimately be a more robust mechanism that will bring rewards to them," says Nepstad, formerly a scientist at the Woods Hole Research Center in Massachusetts

J. ZUCKERMAN/CORBIS

who recently joined the Gordon and Betty Moore Foundation in San Francisco, California, to head its environmental conservation programmes. He is among more than a dozen scientists and economists, including Nobel laureate Joseph Stiglitz of Columbia University, New York, who worked on the proposal, which was organized by the Wentworth Group of Concerned Scientists in Australia.

The number of credits that can be sold would be limited each year, guaranteeing a stable, long-term market. And it requires countries to maintain protected areas, including those for which carbon credits have been sold.

Doug Boucher, a deforestation expert with the Union of Concerned Scientists in Washington, calls the approach "novel" but says it introduces a new set of problems, notably that credits could be sold now when the threat of deforestation is decades down the road. He also questions whether negotiators are open to a radical shift in direction at this point. "There's really some strong momentum now," Boucher says. "That's good, but it does tend to limit the possible options." ■

Jeff Tollefson

SPECIAL REPORT

Where have all the flowers gone?

At least 117 boys were being born for every 100 girls at the beginning of this century in China. **Philip Ball** asks whether Chinese birth rates can be controlled without exacerbating the gender imbalance.

LOU LINWEI/ALAMY

A common female name in China, Laidi, encapsulates one of the country's biggest problems of population management. It means 'a little boy is following', betraying the widespread longing for a son. But tight restrictions on family size have meant that, for many, that son never follows.

The conflict between population policy and the traditional preference for sons is now leaving a legacy of imbalance in the gender ratio, which could foment social tension over the next few decades as the most-affected generation reaches adulthood. "In ten years' time, it will be a real problem," says Therese Hesketh of University College London, who is a specialist on childcare issues in east Asia.

The population challenge for China is very real. The country's last population census in 2000 revealed an increase of 750 million over half a century — more than a doubling. The unsupportable nature of the modern population boom led China's then leader, Deng Xiaoping, to introduce in 1979 the one-child policy, which nearly all citizens were supposed to observe. Financial incentives were provided for compliance; failure to do so drew fines and confiscations of property, and in some cases led to enforced abortions.

"This trend could lead to increased levels of antisocial behaviour and violence."

The one-child policy has had a dramatic impact: the birth rate per woman dropped from 5.4 in 1971 to 1.8 in 2001, and is even lower in urban areas. But the effectiveness of the policy varies. Resistance in rural areas has led to an allowance of a second or even third child, particularly among ethnic minorities. "It's rather like a one-child-and-a-half policy," says Christophe Guilmoto of the Research Institute for Development in Paris. "In many areas, you're entitled to a second birth if the first is a girl." Moreover, he and Isabelle Attané at the National Institute for Demographic Studies, also in Paris, say that "birth control is slipping out of the hands of the regime's cadres, and coercive measures are failing"[1]. The Chinese government has now shifted its emphasis from coercion to voluntarism, with a focus on health issues and education rather than population control per se.

The current Chinese population stands at more than 1 billion — about a fifth of the global total — and is growing at a rate of 8–10 million a year. This growth could impact on China's economic miracle. Niu Wenyuan of the Institute of Policy and Management of the Chinese Academy of Sciences (CAS), Beijing, estimates that nearly a fifth of the newly increased gross domestic product is consumed in feeding the population, although this proportion has been dropping steadily since the late 1980s[2].

Niu, who is a specialist on sustainable development, says that achieving strictly zero population growth is essential. In 1999, a CAS report stated that this target should ideally be reached by 2030. "From 2020 on, the high peaks of total population, aged population and working-age population will come in succession," Niu says. "This will put heavy pressure on China's sustainable development."

The reasons for the Chinese preference for a son are deep-seated, especially in rural areas. The motivation is partly economic: a son may be considered able to work harder in the countryside, or be more likely to get a lucrative job in the city. In part it is about welfare: a son is duty-bound to look after his parents in their old age, whereas a daughter's obligations are transferred to her in-laws when she marries. The importance of a male heir is also a legacy of patriarchal Confucian culture.

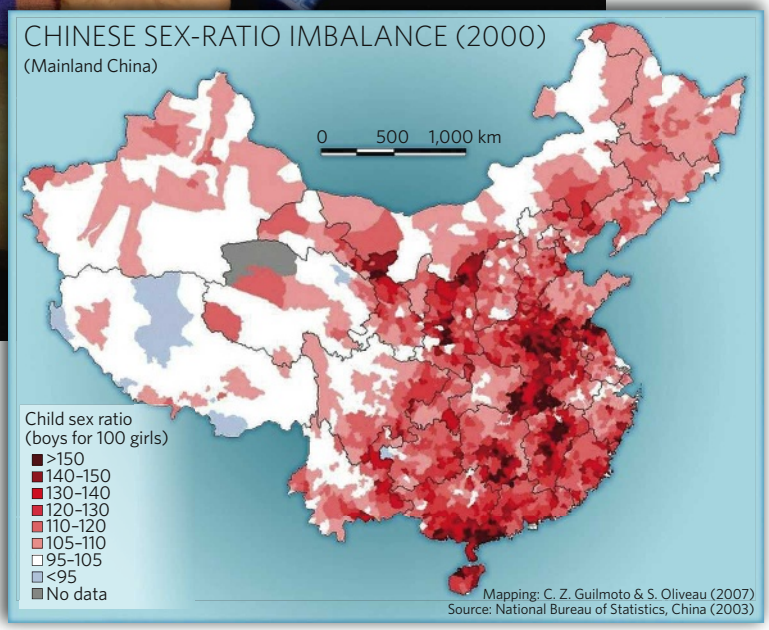
In past decades this has led to the abortion, abandonment and even infanticide of females. In 1982 the average male-to-female ratio at birth in China was 1.07 (as opposed to the normal level of 1.03–1.06); by 2000, various estimates put it at 1.17–1.21. And according to even official figures, the female-to-male infant mortality ratio rose during this period from around 0.95 to 1.46. The timing seems to imply a direct link to the one-child policy, although Guilmoto points out that the sex ratio has also increased in recent times in countries where no such restrictions apply, such as India and South Korea.

Infanticide is now extremely rare. Hesketh says that the higher mortality of baby girls probably stems mostly from a greater reluctance to take sick newborn girls than boys to hospital, where intensive care is very expensive. And modern prenatal and reproductive technologies have created new opportunities to manipulate the outcome of conception.

Missing girls

Hesketh says that the dominant cause of the sex-ratio imbalance is sex-selective abortion, which could account for around 95% of the 'missing females'. "Ultrasound scanning is readily available, even in poor rural areas," she says. Sex-selective abortion is illegal but hard to police. Hesketh feels that stricter enforcement could have a big impact, but new techniques for determining fetal gender could add complications. "The new method is now DNA blood testing, which is much more convenient than anything else," says Guilmoto. "It's not common in China yet, but things may change rapidly."

There is some evidence that couples may be enhancing their chances of having a son by using fertility drugs to increase the likelihood of twins. That is suggested, for example, by a doubling of the number of twins born recently in some hospitals in southern China, where fertility drugs may be available from Hong



CHINA SPECIAL
Read more online at
www.nature.com/news/specials/china/

Kong. Hesketh says that such accounts remain anecdotal, but "I wouldn't be surprised if this were happening".

Yet Guilmoto and Attané have shown that national average figures mask a complex geographical picture[1]. The sex-ratio imbalance in mainland China is generally greater in the countryside than in cities, and it also seems to be concentrated in pockets (see map). "There exist large zones characterized by extreme sex-ratio values bordering areas where recorded values are almost normal," they say. Hesketh says that just three provinces — Henan, Guizhou and Jiangxi — account for much of the elevated female infant mortality. Meanwhile, regions with high proportions of ethnic minorities, such as the far western and northern provinces, tend to have more normal sex ratios. This distribution has remained fairly stable between 1990 and 2000, despite attempts to eliminate sex selection, suggesting that state intervention has had little effect.

In 2006, Hesketh and Zhu Wei Xing of Zhejiang Normal University in China warned that the male–female imbalance could cause serious social tension and disruption in the future[3]. In China there is a strong expectation that young people will marry and have a family, whereas Hesketh predicts that over the next two decades this may be impossible for up to 15% of men. The imbalances are greatest in poorer, rural areas, and because women from this background will be able to 'marry up', it is mostly the poorest men who will find themselves with no marriage prospects. Already, she says, 94% of unmarried people aged between 28 and 49 are male.

"This trend could lead to increased levels of antisocial behaviour and violence," Hesketh says. "When young men congregate, the potential for more organized aggression is likely to increase substantially, and this has worrying implications for organized crime." Already, girls have been abducted to become future brides for families with a son. Female trafficking is on the increase, says Hesketh, and so too is the sex industry. But she cautions that it is very hard to attribute cause and effect when the economic situation in China is changing so fast.

The Chinese government has expressed explicit concerns about the dangers for society and security. In 2004, Li Weixiong, vice-chairman of the population, resources and environment committee of the National Committee of the Chinese People's Political Consultative Conference, said that "serious gender disproportion poses a major threat to the healthy, harmonious and sustainable growth of the nation's population".

Better prospects

The sex ratio at birth should even out. Niu's calculations predict that China's natural population growth rate should drop sufficiently by 2015 for the one-child policy to be adjusted, and that by around 2030 a couple will be able to have two children on average. He therefore predicts that "the sex ratio will drop to 113:100 by 2030, and by around 2050 it is expected to be close to the normal level". Hesketh concurs, saying "I suspect things will settle by 2050".

The Chinese government is eager to change attitudes, and has been running a 'Care For Girls' campaign, which promotes the value of daughters, for the past decade, even in remote rural parts of the country. Legislation has made it easier for girls to inherit, and some provinces have introduced perks and incentives, such as waiving school fees, for daughter-only families. Such efforts may now be bringing results. According to the results of a 2005 micro-census, the sex ratio at birth has already stopped increasing; in fact, Hesketh suspects there is "probably a small downturn occurring". The next national census in 2010 might provide a clearer picture.

"We might have indeed reached a plateau at the national level," says Guilmoto, "although in some inner provinces it might still be increasing." He believes that "China may be entering a new stage in which demand for sons will decline, as it has in South Korea". But he adds that "the excess of male births is bound to seriously disturb the marriage market for decades to come". ■

1. Guilmoto, C. Z. & Attané, I. *Watering the Neighbour's Garden: The Growing Demographic Female Deficit in Asia* (eds Attané, I. & Guilmoto, C. Z.) 109–130 (CICRED, Paris, 2007).
2. Niu, W. Y. *The Overview of China's Sustainable Development* Ch. 10, 259–288 (Science Press, Beijing, 2007).
3. Hesketh, T. & Xing, Z. W. *Proc. Natl Acad. Sci. USA* **103,** 13271–13275 (2006).

See Editorial, page 367, and Books & Arts, page 403.

K. CAMPBELL/GETTY IMAGES

GOT A NEWS TIP?
Send any article ideas for *Nature*'s News section to **newstips@nature.com**

SNAPSHOT

Track record

CHINA

Qi Zhou, a cloning specialist at the Institute of Zoology of the Chinese Academy of Sciences in Beijing, is one of more than 20,000 luminaries selected to carry the Olympic torch on its 137,000-kilometre journey to Beijing for next month's Games.

Zhou had lessons in how to hold the torch for the 200-metre sector that he ran last month. He also had to prepare for possible encounters with protesters, because his leg took place in Xinjiang in northwest China, a region with a large and active minority population of Turkic Uighurs who are striving for independence. "Even though there are some risks, it is worth it to be there," Zhou says. "As a scientist I am so excited to have been chosen as a torch carrier."

It's not his first brush with the Olympics. In 2001, a cow born through a new cloning technique that Zhou helped to develop was named 'Olympic 2008' by his then colleagues at France's national agricultural research organization INRA in Jouy-en-Josas, in celebration of Beijing's selection as host city. ■

David Cyranoski

XIAOJUAN DUAN

Affymetrix in new patents row

The Massachusetts Institute of Technology (MIT) has filed a lawsuit against DNA microarray company Affymetrix, claiming that some of the company's GeneChip technology — widely used for high-throughput genomic analysis — infringes an MIT patent.

Patent disputes are routine in biotechnology, but this one targets a business of growing importance for Affymetrix of Santa Clara, California. The MIT patent, based on work by David Housman, specifically covers "a method for detecting the presence or absence of a single nucleotide polymorphism (SNP) allele in a genomic DNA sample". A popular use of such chips is to search the genome for SNPs associated with disease susceptibility.

"The claims are very broad," says Shaun Rodriguez, a Boston-based analyst for Cowen and Company investment bank. "They cover most of what Affy sells on the genotyping side."

The lawsuit, filed on 1 July in Massachusetts District Court, cites MIT and a patent-holding company called E8 Pharmaceuticals of Cambridge, Massachusetts, as the plaintiffs. Two days later, Affymetrix filed a notice about the lawsuit to the US Securities and Exchange Commission stating: "We believe that the plaintiffs' claims are without merit and will vigorously defend against the claims advanced in the complaint."

Housman's patent was awarded on 9 March 2004. Six months later, Affymetrix filed an application with the US Patent and Trademark Office claiming that the company's 1994 patent application for its basic microarray technology had priority. Last year, the patent office officially disagreed. Affymetrix has continued to produce and sell the disputed GeneChips, and MIT is claiming that the company wilfully infringed the institute's patent.

R. BAER

Affymetrix's GeneChips are big business.

If a jury agrees, the decision could result in hefty royalty payments to MIT. Affymetrix won a similar case four years ago when a jury awarded it 15% royalties from arrays produced by San Diego-based rival Illumina. And there is no word yet about any effect the lawsuit might have on Affymetrix's relationship with the Broad Institute — a genomics institute run by Harvard University and MIT. Some of the company's SNP-based chips use technology developed from this collaboration. ■

Heidi Ledford

WILDLIFE CONSERVATION SOC./EPA/CORBIS

Coral isotopes show quake history

Carbon isotopes trapped for thousands of years in coral skeletons could establish the long-term frequency of major earthquakes in southeast Asia and the South Pacific, and perhaps enable these events to be forecast.

Geoscientists have used corals before to look at earthquake history, by studying the terraced growth patterns that result. A major quake can push up an entire region, thrusting parts of a reef above the low-tide level, killing the exposed coral polyps. The rest of the coral continues to grow, producing a 'hat-brim' pattern that can indicate elevation changes as small as a few centimetres. This phenomenon has allowed scientists to date many earthquakes, including major ones in 1797 and 1833 off Sumatra, Indonesia. But the pattern erodes over time, so it can only be used to identify quakes that occurred within the past few hundred years.

Now, Michael Gagan, a palaeoclimatologist at the Australian National University in Canberra, and his colleagues have come up with a method that relies on a more durable and prevalent record: earthquake-induced shifts in the way that corals store carbon. It could allow researchers to determine the long-term recurrence rate — going back thousands of years — of major quakes in places such as Sumatra, where the Australian tectonic plate is butting up to the plate on which Sumatra rests.

"We can go back further in time and apply it to any reef in an earthquake-prone area," says Gagan, who presented the method on 15 July at the Goldschmidt geochemistry conference in Vancouver, Canada. "I'm being conservative in saying I'm sure we can go back 7,000 years."

The method relies on the ratio of carbon isotopes deposited in the coral by the photosynthetic algae that live symbiotically in it. Algae discriminately process carbon-12 from the water during photosynthesis, and when the light is stronger, for example in the summer, they process more of it. Gagan and his colleagues noticed a similar effect: when an earthquake pushes coral closer to the light, the algae process more carbon-12, leaving a higher proportion of carbon-13 in the surrounding water. The coral, which does not discriminate, therefore absorbs more carbon-13, which can be detected in the coral skeleton.

The sudden shift in isotope ratio — a "big" jump compared with the seasonal average, Gagan says — can be dated precisely. Uranium–thorium isotope dating can pinpoint events to within 1% of the age of the coral, and annual growth bands, like tree-rings, can also be counted. And unlike the existing method, which involves cutting through the coral with a chainsaw, the new method needs only small cores.

Fred Taylor, a geologist at the University of Texas in Austin, who pioneered the 'hat-brim' analysis, says the new method holds promise. "It's a big deal in our little world," says Taylor, who wants to apply Gagan's technique to an earthquake-prone area near the Solomon Islands.

Gagan says the method could also be used to determine whether the recurrence rate for large Sumatran earthquakes has changed over time. A change would indicate that the Australian plate slips erratically below its neighbour, an issue debated by geologists. The method could also be used to distinguish events just a few months or years apart — important if geologists are to learn whether these major quakes come in pairs. For example, geologists suspect that the 1797 earthquake triggered the one in 1833, just as the 2004 earthquake led to another whammy in 2005.

The fault movement needs to have been fairly big — around 30 centimetres — to be detectable and so the method would work best on the big earthquakes, Gagan says. The carbon-isotope signal from smaller quakes could be confused by other events that cause changes in the amount of light a coral receives.

Gagan thinks the method might even help predict quakes. He says that data collected before the 2004 earthquake suggest that there is a change in the subsidence rate of the 'upthrust' region over the five years preceding a quake. If this pattern is borne out in studies of historical quakes, then "maybe it will have some predictive use", he says. ■

Eric Hand

Fusion verdict: misconduct

Nuclear engineer Rusi Taleyarkhan falsified the circumstances of high-profile experiments on bubble fusion, according to a Purdue University report released last week. The report by a Purdue committee that includes scientists from other institutes upheld two charges of research misconduct.

Taleyarkhan's work has been a source of controversy since 2002, when he claimed to have triggered nuclear fusion reactions by passing sound waves into a cell filled with deuterated acetone[1]. His work has been the subject of at least two inquiries by Purdue, which is based in West Lafayette, Indiana. But the latest one was run with oversight from a government agency, the Office of Naval Research (ONR) in Arlington, Virginia, which funded some of the research under question.

"The committee clearly documents that there has never been any successful replication except when Taleyarkhan is present or supervising."

The report finds Taleyarkhan guilty of misconduct for citing a paper by junior researchers in his lab as if their work was an "independent" replication of his own findings. He is also found guilty of adding the name of a student who had not contributed to the paper as an author, apparently in order to counter a reviewer's comment that the replication effort seemed to lack witnesses.

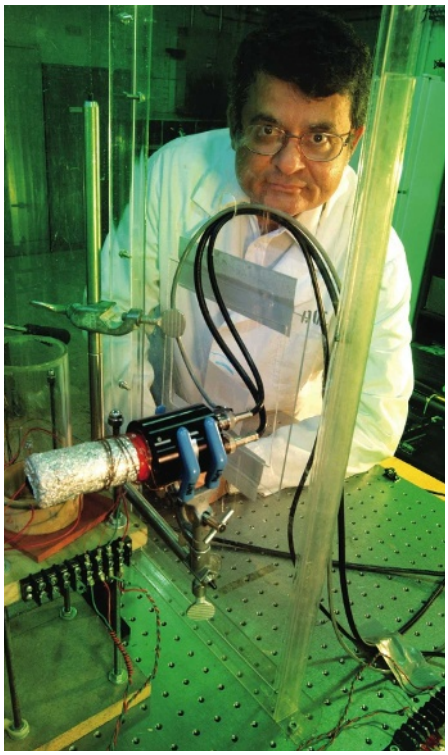
The report stresses that corroborative information should be conveyed honestly, because reproducibility of results by independent experimenters is a crucial component of the scientific method.

Although the report's conclusion echoes concerns expressed[2] by Purdue faculty in 2006, it leaves others unaddressed. The committee of six scientists, chaired by Purdue biochemist Mark Hermodson, notes in its report that it was not sent allegations (from an earlier inquiry) of "intentional data fabrication" relating to the possibility that Taleyarkhan's fusion signal might have come from a radioactive lab source. Two scientists told *Nature* last week that evidence they gave Purdue does not seem to have been considered either. Purdue has not released its charge to the committee; this is a key document that would reveal the questions officials asked investigators to examine.

C. K. Gunsalus of the University of Illinois at Urbana-Champaign, who is an attorney and an expert in research misconduct, says that it is good practice for a university to turn over all of its material to an investigation panel and to set a broad charge. "Their findings of fact are rigorous, but the committee clearly documents that there's never been any successful replication except when [Taleyarkhan] is present or supervising, and they don't explore the implications of that," she says.

In a letter released by Purdue, the ONR inspector general Holly Adams calls Purdue's investigation "prompt, thorough and objective", but says she is still waiting to hear what corrective action the university will take. Unusually, Purdue floated news of the misconduct finding while Taleyarkhan still has 30 days to appeal. Taleyarkhan did not respond to *Nature*'s request for an interview, but in a statement released on 18 July, his attorney, John Lewis of Lewis and Wilkins in Indianapolis, Indiana, said that all charges except two had been "resoundingly" resolved in Taleyarkhan's favour. ■

Eugenie Samuel Reich



L. FREENY/US DEPT. ENERGY

Rusi Taleyarkhan: found guilty of misconduct.

1. Taleyarkhan, R. P. *et al. Science* **295,** 1868–1873 (2002).
2. *Nature* doi:10/1038/news060306-2 (2006).

SIDELINES

ON THE RECORD

"When you're eating your sandwiches on the lawn you don't expect to find something that takes you by surprise."

Entomologist Max Barclay, a curator at London's Natural History Museum, on finding a mysterious insect in the museum garden. The tiny bug is not matched by anything in the museum's collection and might be a new species.

NHM

SCORECARD

Drinking
Loud music makes people drink more and faster, researchers report after visiting bars over three Saturday nights and messing with the sound levels. Sidelines is penning its grant proposal now.

Eating
New York City rules to curb obesity now force restaurants to post calorie information on their menus. The good news is that those too busy for a long dinner can satisfy their daily calorie intake with a couple of burgers.

SCORECARD (AGAIN)

Touchy-feely physicists
Hold a boson to your bosom. Physicists in need of a cuddle can now grab Subatomic Particle Plush Toys.

THE PARTICLE ZOO

Touchy-feely PC users
An IT analyst says the computer mouse will be extinct within five years, replaced by facial recognition software and other devices.

WORDWATCH

Make-make
The new name for the third dwarf planet discovered, which for the past three years has had only a licence-plate number and was saddled with the humiliating moniker 'Easterbunny' by astronomers wishing to reference it.

Sources: BBC, Reuters, AP, The Particle Zoo, www.mikebrownsplanets.com

Roche bids for remaining Genentech stake

The Swiss pharmaceutical giant Roche has offered nearly US$44 billion to acquire the 44% of biotechnology jewel Genentech that it doesn't already own.

The offer immediately raised questions about how independent the highly successful, South San Francisco-based biotechnology firm would remain. Severin Schwan, the chief executive of Roche, says that his company "will take the necessary steps to nurture Genentech's innovative and unique science-driven culture".

Genentech has a market value of some $86 billion and has generated biotech blockbusters such as Avastin (bevacizumab) for colon cancer and Herceptin (trastuzumab) for breast cancer. Roche acquired its majority stake in Genentech in 1990. The Swiss firm expects to complete the transaction "as soon as possible", subject to approval from the holders of the majority of Genentech shares not already owned by Roche.

For a longer version of this story see http://tinyurl.com/5pztch

German public–private partnership breaks ground

For the first time in its 60-year history, Germany's Max Planck Society (MPS) is setting up a new institute using private money — €200 million (US$317 million) of it. On 15 July, the twin brothers who founded the generic drugs company Hexal signed an agreement with the MPS to establish an institute for cognitive neuroscience in Frankfurt. The institute is set to start work by the end of this year.

The Ernst Strüngmann Institute is named after the father of Hexal founders Andreas and Thomas Strüngmann. It will have the same scientific independence as the 80 other MPS institutes, but unlike them its finances will be overseen by a separate board, on which Andreas Strüngmann will sit.

"Public–private research is a well established concept in the United States, but not yet in Germany," says Wolf Singer, director of the Max Planck Institute for Brain Research in Frankfurt and provisional head of the new institute. "I hope this model will become common practice."

Clinical trialists less likely to seek grant renewals

Clinical grant proposals receive poorer scores than their basic science counterparts from reviewers at the US National Institutes of Health (NIH), partly because clinical trialists are less likely to seek grant renewals.

A study published this month (M. R. Martin *et al. Am. J. Med.* 121, 637–641; 2008) examines review outcomes of almost 63,000 basic-science grant applications and more than 30,000 clinical grant applications reviewed between 2000 and 2004. It found that clinical applications were scored less favourably, which agrees with previous research.

About half of the difference was due to the failure, by some 15% of clinical proposals, to adequately address human-subject protection, the researchers say. The remainder was due to fewer clinical trialists (20%) competing to have their grant applications renewed compared with their basic science colleagues (28%).

This made a difference because the overall success rate of grant applications climbed with resubmissions: by the second round of resubmissions, differences between clinical and non-clinical application success rates had evaporated.

Ontario acts to protect its boreal forests

J. WROBEL/ALAMY

Some 225,000 square kilometres of Ontario's boreal forest have been set aside in the biggest conservation initiative in Canada's history. But mining and logging will still be permitted, with strict regulations, in the region.

Ontario premier Dalton McGuinty introduced the plan last week. Boreal forests make up half the province's land area, and are estimated to sequester 12.5 million tonnes of carbon dioxide each year.

On 18 July, Ontario also announced it was joining the Western Climate Initiative, a regional group with goals to cut greenhouse emissions. Seven US states, plus Quebec, Manitoba and British Columbia, are the other members.

J. SULLIVAN/GETTY IMAGES

Digital libraries often rely on English text.

Google Books expands its non-English resources

The municipal library of Lyon, which holds France's second largest collection of books, going back centuries, has become the first French library to join forces with Google.

The Google Books Library Project will digitize at no cost half a million of the library's 1.3 million texts whose copyright has expired, in return for an exclusive 25-year licence to the digital texts.

Lyon's decision runs counter to criticism of the Google library project by French historian Jean-Noël Jeanneney. In 2005, while president of the National Library of France in Paris, Jeanneney argued that Google's domination of a universal digital library would result in an over-reliance on English texts.

The National Library itself has plans to digitize 300,000 of its texts, and a Europe-wide digital library of cultural and scientific heritage is under construction.

US Senate approves $48 billion global AIDS funding

Following a 16 July vote by the US Senate, the House of Representatives was expected this week to approve $48 billion in global spending on AIDS, malaria and tuberculosis between 2009 and 2013.

The House action would clear the way for President George W. Bush to sign into law the renewal of a respected five-year-old programme, originally dubbed the President's Emergency Plan for AIDS Relief.

The House was not expected to oppose a Senate-adopted provision that would reverse a 21-year-old ban on HIV-positive visitors entering the United States.

Separately, the US National Institute of Allergy and Infectious Diseases said last week it is cancelling its planned large-scale trial of an HIV vaccine, citing disappointing results from a trial of a related Merck vaccine last autumn.

For a longer version of this story see http://tinyurl.com/673f29

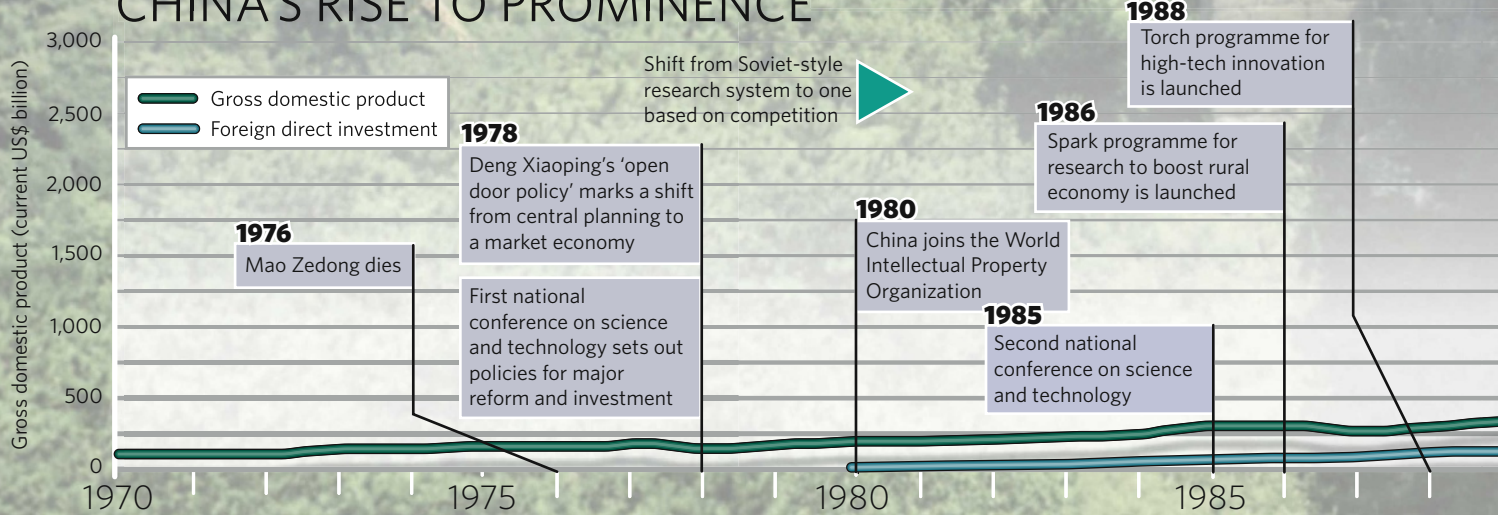
THE GREAT CONTENDER

CHINA

China's performance has been remarkable in any number of fields. **Declan Butler** charts the country's scientific and economic growth.

The map (above) and cartogram (left) reveal mainland China's regional disparities. Estimates of population density for 2005 show a concentration in the east, especially along the coast. This matches an explosion in urbanization: more than 40% of the population now lives in cities, compared with 16% in 1960. In the cartogram the area of each region is adjusted according to its share of gross national product (GNP) per capita. Economic activity is densely concentrated in municipalities such as Beijing and Shanghai and in the provinces of the eastern seaboard. Cartogram created with ScapeToad (http://chorogram.choros.ch/scapetoad).

CHINA'S RISE TO PROMINENCE



In the Olympics of scientific and technological performance, China has surged forward from the pack trailing at the back to overtake many of the long-standing pacesetters. Its impetus is such that it is surely only a matter of time before the country secures a place on the podium.

China began going for gold with Deng Xiaoping's 1978 'open door policy', the first round of post-Maoist reform. The subsequent shift towards a market economy increasingly open to outside investment has sent almost every financial indicator shooting off the graph, with the economy growing consistently by around 10% annually. In foreign direct investment, China now far outperforms Japan or South Korea, and enjoys levels similar to those of the United Kingdom or France. High-technology exports grew from 6% of all manufactured exports in 1992 to 30% in 2006. Its overall gross domestic product (GDP) was 11th in the world in 1980, just behind Mexico; in 2006 it was 4th, behind Germany.

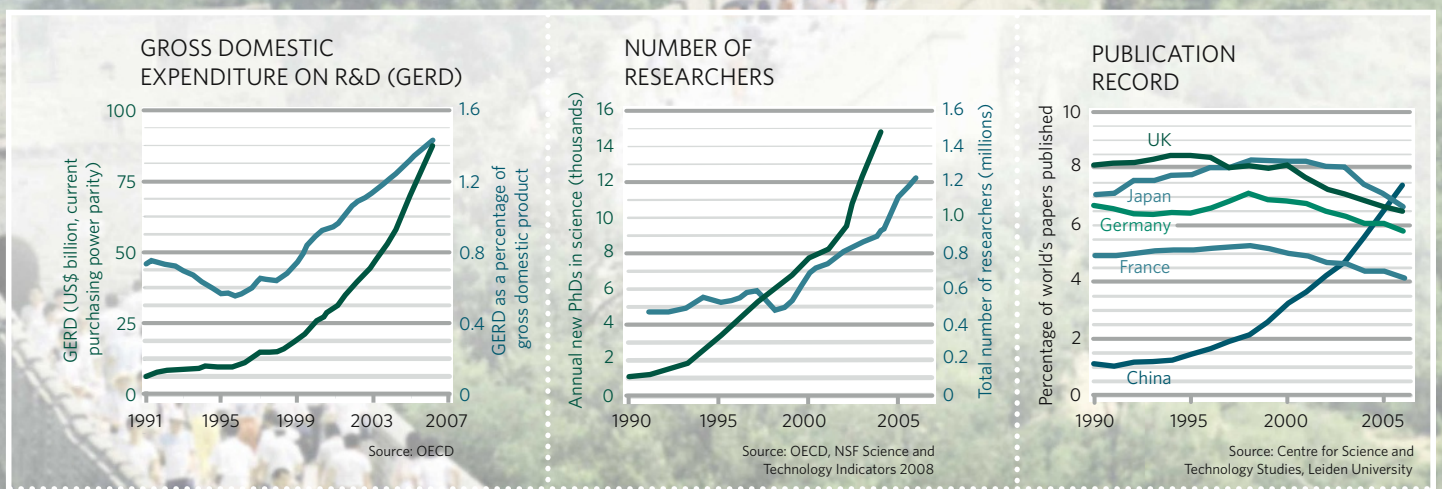
These riches have lifted hundreds of millions of people out of poverty, although rural areas have benefited less than the eastern seaboard and megalopolises such as Beijing and Shanghai, where the wealth is concentrated (see map). Research spending is more evenly spread than wealth, yet the majority of it also finds its way to these eastern regions.

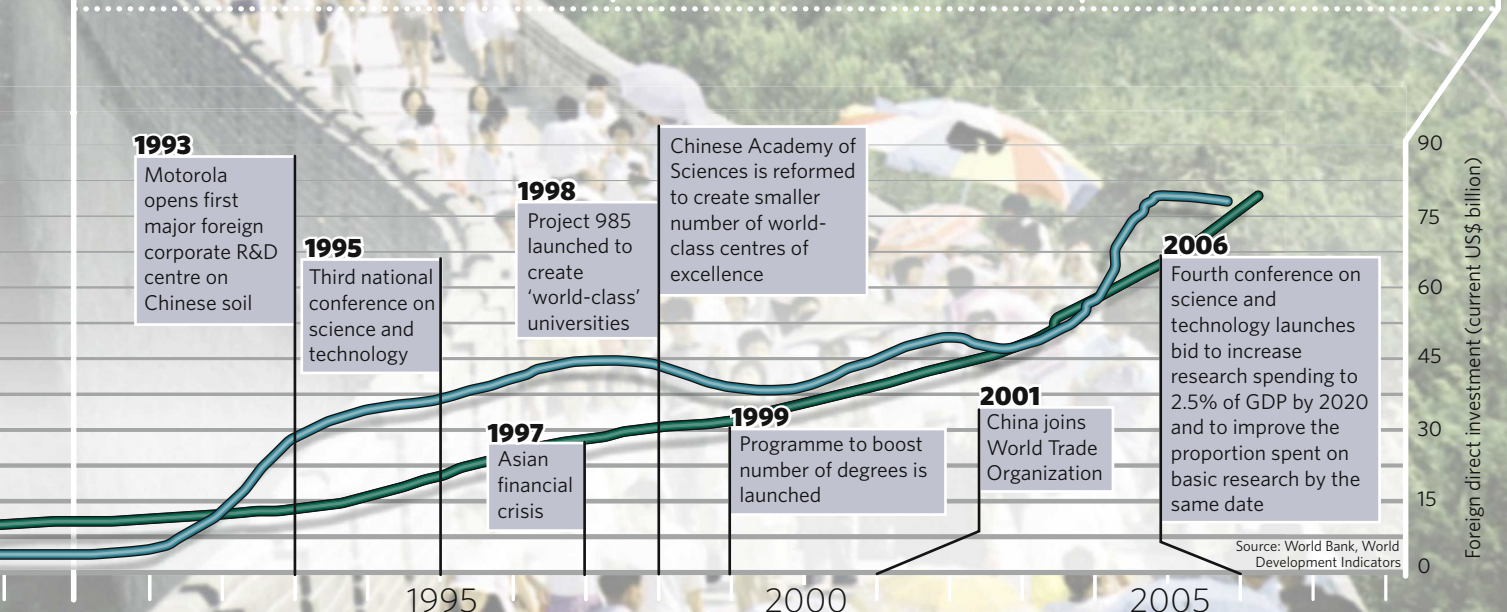
Over the past two decades, spending on research has grown by almost 20% each year, increasing from US$7.5 billion (calculated on a basis of purchasing power parity) in 1991 to almost US$90 billion in 2006. The proportion of total research money spent on basic research, at just 5%, is far lower than in most advanced economies, where rates reach up to 20%. China is seeking to boost basic research and innovation to secure future economic growth — 'indigenous innovation' is the theme of the national science and technology plan for 2006–20.

The number of students taking science or engineering degrees in China each year climbed from 115,000 in 1995 to more than 672,000 in 2004, putting the country ahead of the United States and Japan; about two-thirds of the Chinese degrees were in engineering. In 2007, Chinese scientists accounted for 32,000, or almost one-quarter, of the 142,000 foreign students receiving PhDs in the United States, more than any other country except India, which accounted for one-third.

China's share of the world's published scientific articles soared from 0.2% in 1980, to 7.4% in 2006, when it overtook Japan for the first time. Productivity in this respect is not yet matched by quality. The citation impact score for Chinese research — a measure defined so that the average for the world is 1.0 — remained stagnant at about 0.35 for decades. In the decade from 1995, it doubled to 0.73, and in materials science and nanotechnology, two of China's specializations, its score is approaching 1.0. ■

IMAGE SOURCE/GETTY IMAGES





VISIONS OF CHINA

Can the Chinese government meet its ambitious targets on space, the environment, research, energy and health? **David Cyranoski** takes a look at China today and what it hopes to be tomorrow.

China's post-Mao leadership has never been under so much pressure. At home, it is trying to appease 1.3 billion increasingly politicized citizens. Abroad, other world powers are demanding more responsibility in environmental, economic and geopolitical issues. The government has its own agenda, and is increasingly calling on science and technology to help. It is clear that expectations are high as the country moves ever closer to the centre of the world's stage. What is not known is whether China is capable of meeting others' expectations — or its own bold targets.

Space: slowly but surely

In September 2005, spurred on by President George W. Bush, NASA officials said they could put a man back on the Moon by 2018. Two months later, Ouyang Ziyuan, the head of China's lunar exploration programme, was widely quoted as saying China would do the same by 2017. The media portrayed a space race, with the two world powers matching each other stride for astronomical stride.

The reality is more prosaic: China has never officially announced plans to land on the Moon. The report may have been a misquote, a mistake or a deliberate leak intended to goad its competitors, but it does not fit with China's conservative approach to space exploration, which proceeds by small, realistic goals. "China is not in a conventional race," says Eric Hagt, China programme director at the World Security Institute in Washington DC. "If it is, it is more like the hare versus the tortoise."

As in other fields of endeavour, China's objectives for space are laid out in five-year plans; the tenth plan covered 2001–05. The country's biggest accomplishment was to send a human into orbit in 2003, becoming only the third country in the world to do so. China also delivered on its promise of improving remote-sensing capabilities with the launch of the Haiyang-1 ocean-imagery satellite, three Earth-observing satellites and upgrades to disaster-monitoring satellites. All in all it was "a fairly impressive job", says Hagt.

China moves into its eleventh five-year plan with confidence. Its goals this time include a Moon orbiter, which was successfully sent up in 2007, and next year the launch of Shijian-10, a recoverable satellite. There are also plans for the launch of a joint unmanned Mars mission with Russia next year, the country's first astronomy satellite, for black-hole research, by 2010, and a solar-flare observatory with France a year later. The nation's incremental plans serve many functions, including the development of surveillance satellites for national defence, for the creation of spin-off technologies and for disaster management: satellite images played a pivotal role in monitoring lakes formed in the aftermath of the recent Sichuan earthquake. "It is less about vision and more about key national development goals," says Hagt.

The US space programme, by comparison, is all about vision. Bush's 2004 Vision for Space Exploration "was always more fantasy than actual plan", says Joan Johnson-Freese, a specialist in national security at the Naval War College in Newport, Rhode Island. The United States is still ahead of China in space technology. But some current US goals, such as those to develop and start testing Orion, a new spacecraft, by 2008 and to complete the International Space Station by 2010, are already in doubt. More ambitious ones, to return to the Moon and send people to Mars, face an uncertain political and financial future. Although China's US$500-million annual space budget is dwarfed by NASA's $16-billion one, its goal-setting seems more realistic.

... OR BLEAK OUTLOOK

ILLUSTRATIONS BY DAVID PARKINS

As China hits its launch targets, will it also honour its commitment to peaceful use of space? Almost all space technology is dual-use. Debate flared last year when China zapped one of its old satellites out of the sky with an anti-satellite weapon, and Hagt says "the gap in its rhetoric and action is opening up". It might be fear of space wars that has the United States acting like China's jittery competitor. "[China] has the political will to plod slowly, incrementally," says Johnson-Freese. After all, everyone knows who won the race between the hare and the tortoise.

Environment: shades of green

There are many reasons why China wants so badly to be green — pandas, the Olympics and healthy citizens to name a few. Despite significant efforts over the past five years, the country's coal-driven manufacturing boom has painted its reputation in shades of grey. According to the World Bank, 20 of the world's 30 most-polluted cities are in China — Beijing and Shanghai among them. "Year after year, the country ranks at or near the top of world charts in terms of land degradation, air pollution, and water pollution and scarcity," says Elizabeth Economy, director of Asia studies at the Council on Foreign Relations in New York.

There is no doubt that China's upper levels of government are committed to environmental improvement, says John Mackinnon, a veteran of biodiversity studies in China currently working on the European Union–China Biodiversity Programme. The problem is that government agencies often work against each other and come up against contradictory laws. The forestry agency might try to protect wetland nature reserves, for example, but the water-resources ministry will give permission to pump water. Furthermore, Mackinnon says, the government rarely involves active scientists in the decision-making process. "They don't want to pay outsiders and they don't think they need any help."

Ahead of the Olympics, China has been brandishing its report card on 'blue sky days' in Beijing, when measures of sulphur dioxide, nitrogen oxides and particulate matter fall below 100 on a 500-point scale — a measure that would still be considered heavy pollution in other countries. There were only 100 such days in 1998. But after pledging 245 days last year, the government announced that it had succeeded with 246, and is on target for this year's goal of 256. The success is credited to 120 billion yuan (US$17.6 billion) spent on air-pollution control measures since 1998 and, over the past two years, an expansion of public transport and relocation of factories. But a report by Steve Andrews, an independent environmental consultant based in Washington DC, and once a fellow at the Natural Resources Defense Council in Beijing, said the clearing skies actually resulted from a repositioning of the monitors: the exclusion of two in more polluted areas and addition of three in cleaner ones. "Instead of meeting standards, China sometimes restates standards to meet reality," says Drew Thompson, director of China studies at the Nixon Center, a think-tank in Washington DC.

On the panda front, China is also making commitments, signing up to the Convention on Biological Diversity in 1992 and the Ramsar Convention on Wetlands in the same year. Thanks to more than a decade of efforts, 15% of the country is now taken up by some 2,700 nature reserves. Regulations are difficult to enforce, though: animals are hunted and medical herbs are collected. "We call these 'paper' nature reserves," says Yan Xie, a biodiversity conservation expert at the Chinese Academy of Sciences (CAS) Institute of Zoology in Beijing. "They just can't stop people from using the land." Giant pandas (*Ailuropoda melanoleuca*) are among those species that have made comebacks; others, such as the Yangtze River dolphins (*Lipotes vexillifer*), are now thought to be extinct in China.

There are increasingly strident calls for change from the internal environment ministry, non-governmental organizations and neighbouring countries affected by Chinese pollution. Economy says that real change will require "Chinese officials and businesses opening themselves and their environmental practices to greater scrutiny and accountability by the Chinese people. This is not yet something that China's leaders seem willing to do."

Research: self-sufficiency

Fears of environmental destruction and energy depletion have accompanied every stride of China's economic growth. The proposed solution by its leaders is to wean the country off energy-intensive, low-end

manufacturing, such as steel, cement and paper, and to create by 2020 an "innovation-driven society".

The specifics were spelt out in the national medium- and long-term programme for science and technology in 2006. The document lists 11 predictable "key industries", including energy, mining, the environment and information technology, and ten basic-research programmes, including protein studies and nanotechnology. The country aims to grow the amount of gross domestic product (GDP) devoted to science and technology from 1.4% as it stood in 2006, to 2.5% by 2020. Such a jump would place China alongside the biggest investors: in 2006 the United Kingdom invested 1.8% of GDP, Germany 2.5%, the United States 2.6% and Japan 3.3%.

Also by 2020, the science and technology programme says, the percentage of technical innovation coming from overseas should drop from around 60% to 30%. China has pledged to become one of the top five countries in its number of new patents and the impact of basic-science research papers, and to boost the number of university graduates.

These are uncomfortable goals for China because, unlike space research, they are more difficult to mandate from the top down: industrial outlays accounted for more than two-thirds of research-and-development funding in 2006. And the tax breaks and other funding mechanisms that the government is using to encourage home-grown technology are taking time to sink in: a 2007 survey by the CAS Institute of Policy and Management in Beijing found that most companies were not even aware of the support measures. It may be more important to remove obstacles, such as the red tape surrounding drug approvals that can be two to three times more time consuming than in the United States or Europe. "It's not more or less regulation — it's better regulation," says Lan Xue, a science-policy specialist at Tsinghua University, currently on sabbatical in a World Bank post in Washington DC.

The number of scientific papers published in China has already doubled in the past five years to 80,000, according to a survey of Chinese research published this month by *Science Watch*, a newsletter tracking trends in scientific activity. This pushes it past Japan and places it second only to the United States. But the average impact of those articles, judged by the number of times each was cited, was below average even in China's strongest fields of materials science, physics and chemistry. "We need to increase the quality," says Rongping Mu, a policy expert at the Institute of Policy and Management, adding that the same is true for patents. "Top five in numbers is not so difficult. But will they be active, useful patents?"

Xue says that the country also struggles to capitalize on those patents because there is no viable venture-capital market. China has established 54 national high-technology science parks. Beijing's Zhongguancun industrial zone, the first such park when it was established in 1988 near Peking University and Tsinghua University, has grown rapidly and turned 400 billion yuan of revenue in 2008. But the park has been criticized for its low output of patents (6,000 applications last year), and the majority of science parks elsewhere have little to show. "The only real importance [of science parks] is as a physical expression of a policy to support high technology," says Stuart Macdonald, a technology-management specialist at the University of Sheffield, UK. "Those in China are no exception."

Innovation is a time-consuming process, and this is frustrating for a country that is hooked on speedy development. "You need a certain culture to support an innovative business environment," says Xue, "and those things change much more slowly."

AP PHOTO

G. NIU/GETTY IMAGES

(Top) Shenzhou VI is part of China's progressive space programme; (bottom) Beijing pledged to reduce its smog in time for the Olympics.

Energy: renewed efforts

At the Renewables 2004 conference in Bonn, Germany, a Chinese delegation gave the first notice of the country's commitment to alternative sources of energy. Speaking to the hundreds of energy experts who packed the meeting room, the delegates promised that 16% of China's energy would be from renewable sources by 2020. It would more than double renewable energy's proportion of a rapidly growing energy demand, most of the rest coming from coal.

China's galloping economic growth is stressing the environment and global fuel supplies; it needs to find new sources of energy and to make the most of those it has. But until that point, observers say that laws and resolutions promoting this goal were not enforced. So the announcement at the meeting made some wonder. "I was surprised that China would announce bold targets beyond what any other country had, except perhaps the European Union (EU)," says Eric Martinot, a renewable-energy specialist who was partly persuaded by that statement to start a three-year stint at the Tsinghua BP Clean Energy Research and Education Centre in Beijing.

China's overall target — since chiselled down to 15% — falls short of the EU's, which aims to supply 20% of energy from renewables by 2020, but it is still ambitious considering the country's massive energy demands. China announced its Renewable Energy Law in February 2005, and in September 2007 released a development plan for renewable energy that is broken down into sectors, with hydropower accounting for the lion's share but including biomass, wind, solar and geothermal.

The objectives are being met with policy measures such as a slowed approval of new coal plants (requiring a year or more longer) and fast-track approval of new wind stations. Billboards with photos of windmills, solar cells and grassland for biomass garner public support. And with an investment of 82 billion yuan in 2007, China became the world's second-largest investor in new renewable capacity, behind Germany and ahead of the United States.

Much of this push came before the official targets were penned, with the result that, even before the September 2007 plan was formally

announced, the target of 5 gigawatts of wind energy capacity by 2010 had already been achieved . Wind has been the greatest success. Capacity has doubled every year for the past three years, pushing China's global wind capacity ranking to fifth in 2007.

Other renewable-energy targets will be more challenging. Getting the stated 30 gigawatts of energy from biomass by 2020 would require 1,000 power plants, but only a handful have opened in the past two years. Overall, though, China looks likely to meet its goals on renewable energy. "There's no doubt the target will be met and probably exceeded," says Martinot.

Renewable energy alone cannot tackle the country's environmental problems. China is also pushing clean-coal technology (see page 388), nuclear energy and strict improvements in energy efficiency. Last year, China tied local officials' promotions to their ability to meet energy-efficiency and emissions-reduction goals. "Now these goals are given the same emphasis as GDP and population-control goals," says Jiang Kejun Jiang, a researcher at the National Development and Reform Commission's Energy Research Institute in Beijing. And those are policies with proven track records.

G. BAKER AP PHOTO

G. BAKER/AP PHOTO

(Top) China aims to generate 15% of energy from renewable sources by 2020; (bottom) health policies were revamped after the SARS outbreak.

Disease: bill of health

In March 2003, as Hans Bekedam listened to Zhu Rongji's farewell speech, he waited and waited — until Zhu was on page 34 of a 42-page speech — for China's retiring premier to make the first mention of health-related matters. Just over half a year earlier, Bekedam had become Beijing representative for the World Health Organization (WHO). Despite simmering problems with heart disease, smoking, HIV, hepatitis and a host of preventable and non-preventable diseases, all the leadership laboured on about was economics. "No one talked about health," says Bekedam.

Sudden acute respiratory syndrome (SARS) changed everything. During the outbreak in early 2003, local surveillance systems failed. More than 5,300 mainland Chinese were infected and 349 died — numbers much higher than they should have been because of the delay. China was heavily criticized internationally. The health minister resigned.

Clearly, China's health system wasn't up to the task, in large part due to a lack of government support. Hospitals were getting only 10% of their funding from the government and relied on patient payments for the rest, yet the vast majority of needy patients lacked health insurance. In such a system, surveillance and treatment — of SARS or anything else — is difficult. "SARS made clear that the government needed to take responsibility for the safety of its people. It can't leave it to the market," says Bekedam. "After SARS they started spending."

By March 2004, the leadership's rhetoric had taken on a different tone. Speaking at the National People's Congress, Zhu's successor Wen Jiabao put "the victory against SARS" top of his list of achievements. "We need to maintain a high degree of vigilance and take firm and effective measures to control SARS, AIDS, schistosomiasis and other serious communicable diseases," he said, promising a system for disease prevention and control in three years.

Since the 1990s, the government had been promising to establish a strategic plan for HIV/AIDS prevention and control, but the disease had yet to be taken seriously. Yet with strong shows of support at the top, things started to change. In 2004, China greatly increased surveillance of high-risk groups and intensified screening of potential blood donors. The number of newly reported cases of HIV infection more than doubled to 47,606, and by 2007 an estimated 700,000 were found to be infected — a grave epidemic, albeit one that is not on the scale of that in Africa.

In early 2006, a five-year Action Plan for Reducing and Preventing the Spread of HIV/AIDS was released. Thanks to government-supported programmes and investment, the number of patients receiving antiretroviral treatment jumped from 5,000 in 2003 to 39,000 in 2007, and many needle-exchange programmes started up. China still has much to work on. A United Nations report covering HIV/AIDS in China from 2006 to 2007 found that access to treatments is sorely lacking, particularly for those in remote areas. Access of the rural poor to health care "is greatly limited, given China's economic status", says Ray Yip, who was assigned to China for nearly a decade by the US Centers for Disease Control and Prevention in Atlanta, Georgia, and is currently with the Bill & Melinda Gates Foundation in Beijing.

Such assessments call into question the country's stated aim to give 80% of HIV sufferers access to antiretroviral treatment by 2010, although Bekedem says that there is genuine commitment. If China can prove its mettle with HIV, it will give the leadership more legitimacy as it approaches an even greater goal — universal access to health care by 2020.

The rest of the world has an interest in strengthening China's health care, anticipating as it does that another disease such as SARS could emerge there undetected. The country now has impressive surveillance policies in place, says Yip, including requirements that local doctors list all cases of 31 diseases into a national database. After the recent earthquake, the government issued special phones keyed with lists of symptoms, and within two weeks they had a surveillance system set up for emerging diseases, although it found none. Yip compares China's response with that of the United States after the Hurricane Katrina disaster: "If this were the Olympics, China would have got a 9.8. The United States would have got a 2.3, if that." ■

David Cyranoski is *Nature*'s Asia Pacific correspondent.

See Editorial, page 367; News Feature, page 388; and Commentary, page 398.

STOKING THE FIRE

China burns more coal than any other country; how it does so in the future will determine our planet's climate. **Jeff Tollefson** reports from Beijing.

CHINA

Huang Bin, a 30-year-old engineer, is surveying the scene at one of China's showcase energy projects: a retrofit that will make the Gaobeidian coal-fired power plant in Beijing burn just a little bit cleaner. Three engineers in red hard hats pore over a blueprint, their fingers tracing lines on paper splayed across a steel tank. Two workers adjust a valve nearby, one of hundreds on a two-storey platform erected alongside two 30-metre-high vessels that will house the chemical reactions at the heart of the project. Sparks fly as welders connect pipes; the buzz of grinders comes and goes. The ground was broken on this project just three months ago, and even an outside observer can tell that there is plenty still to do. But no one seems to doubt that the world's latest carbon-capture pilot plant will be finished in three weeks' time. "Chinese speed," Huang says with a smile.

That was in late June. Last week, as planned, the new unit began stripping carbon dioxide out of a small stream of exhaust from the plant, a high-efficiency, 1,065-megawatt monster that churns out 10% of Beijing's power and one-third of its hot-water heat. The Huaneng Group, the government-owned company that runs the plant, plans to collect less than 1% of the CO_2 emitted here, ultimately to provide some of the fizz in locally made carbonated drinks. It is a modest goal, but for China the project is a gesture of goodwill, a tentative step into the kinds of technologies needed to decarbonize an economy that derives more than two-thirds of its energy from coal.

For years, China has lagged behind the West in researching ways to burn coal more cleanly, but that is beginning to change. Huang and his colleagues are coming of age in an era in which the Chinese learn by doing, and what they are doing today is advanced coal technology. The total time for the Gaobeidian retrofit from announcement through design and commissioning was nine months.

Chinese speed has raised entire cities and built modern highways, all while providing at least basic energy services to most of its 1.3 billion people. It has also frightened a world already alarmed by global warming. The planet's most populous nation has added some 170 gigawatts of coal-fired power capacity in the past two years alone — more than double Britain's entire electricity-generating capacity, installed over a century — and has overtaken the United States as the world's largest emitter of greenhouse gases.

Yet China's single-minded determination to get things done, if properly harnessed, could drive down costs and commercialize advanced coal technologies that have languished in labs and boardrooms in the West. In many ways, China has already positioned itself at the forefront of coal technology, but 'advanced' does not necessarily mean clean. Climate-friendly technologies would enable companies to capture and pump CO_2 underground, eliminating most of the emissions from coal. By contrast, even new technologies for converting coal into transportation fuels without carbon capture might increase China's reliance on coal, as well as its emissions.

"It's relatively easy for me to imagine the Chinese will get way out in front of us in the United States and Europe," says Kelly Sims Gallagher, an expert in China energy at Harvard University in Cambridge, Massachusetts. "The Chinese are committed to installing advanced technology. The question right now is which technology it will be."

So far, China's industrial revolution resembles a compressed version of that experienced in the West, with all the associated environmental problems and resource limitations. Evidence suggests that solutions, too, may come in rapid-fire fashion. An oft-cited statistic is that the Chinese bring a new power plant on line every week or two; less appreciated is that today's power plants generally employ state-of-the-art combustion technology, whereas older, less-efficient plants are being shut down.

The main goal is to save coal. China's coal reserves rank as the world's third largest; the country mined and then devoured some 2.5 billion tonnes of coal last year, more than double the tonnage of the next-largest user, the United States. Still, the mining industry has struggled to meet demand, and imports are on the rise.

Efficiency drive

The government has also made energy efficiency its de facto climate policy (see 'Kicking the coal habit', overleaf), beginning with an ambitious effort to cut energy intensity (the amount of energy consumed per unit of gross domestic product) by 20% from 2006 to 2010. The emphasis is on the manufacturing industries, such as iron, steel and cement, which consume 68% of the nation's electricity and even more of its overall energy. It's not clear whether China will meet that goal on time. For many observers, though, what makes the policy real is the fact that national communist leaders now grade local officials according to their progress on energy efficiency.

The government is also taking aim at conventional pollutants from coal-fired power plants, hoping to curb acid rain and the dense smog that envelops many of China's cities. Roughly half of China's power plants are now equipped with 'scrubbers' for sulphur dioxide emissions. Most of these have been installed since 2006, and there are more to come. "China now has more scrubber capacity than all of the rest of the world put together," says Robert Williams, a senior scientist at the Princeton Environmental Institute in New Jersey. Nitrogen oxides are likely to be next on the clean-up list.

According to the official government line, such efforts are intended to create a wealthier and more 'harmonious' society. At the same time, leaders are under pressure from an increasingly large and vocal middle class that aspires to a cleaner, more prosperous lifestyle. Also telling is that the government has acknowledged in public documents the cold economics of pollution-related deaths and disease. Pollution is likely to slash the country's gross domestic product by anything from 2% to an eye-popping 18% by 2020, depending on how successful the clean-up initiatives are, according to Ming Lei, an environmental economist at Peking University.

Those estimates do not include the potentially enormous effects of climate change. Both politicians and scientists foresee huge problems with increased floods, dwindling crop yields, and less freshwater run-off as Himalayan glaciers recede (see page 393). But only a new carbon economy or a regulatory directive from the government — probably preceded by some kind of international climate agreement — is going to change the status quo, as businesses currently have no incentive to curb CO_2 emissions. "Unless you tell them this will make them money, then they say 'no'," says Zhang Hai, an engineering professor at Tsinghua University in Beijing. "Nobody is a volunteer."

> **"When the Chinese government says it is going to do something, it will do it."** — Lu Xuedu

Fifty kilometres southeast of Beijing, in the industrial city of Langfang, China's energy economics are on display at the new headquarters and research facilities of ENN. This independent company with global ambitions is now betting big on technologies for converting coal into a substitute for diesel fuel. A visitor is driven up to the new, six-storey office building in a black company Audi, complete with tinted windows. Outside, bulldozers are clearing land for three new labs; already standing are pilot plants for coal gasification, biofuels and solar projects, as well as a solar-cell manufacturing plant. All of this has been achieved in the past year, during which ENN has hired some 4,000 employees, boosting its workforce by 20%. The facility has its own hotel, restaurant and golf course. Next up: a university.

ENN began as a rental-car company in 1989 and made its money as a distributor of natural gas and other fuels. It is now pursuing coal gasification, an old technology that glimmers whenever petroleum supplies seem threatened, and is a leader in the new wave of interest in underground gasification. At its simplest, the technique involves drilling a well, igniting the coal within it and adding oxygen; another well sucks out the resulting 'syngas', a mixture of mostly hydrogen, carbon monoxide and CO_2. The syngas can then be condensed to make liquid fuel or chemicals. Most of the early research projects in this area have run their course, although one commercial project has been operating in Uzbekistan for more than four decades. Other companies are planning projects in the United States, Canada and beyond. ENN says it has been operating two pilot projects in Inner Mongolia since last year and is now developing a commercial-scale facility.

Gan Zhongxue, ENN's chief scientist, readily admits that his company is several years behind some of the most advanced Western companies pursuing gasification technologies. ENN purchased its first gasifier from US-based General Electric, one of several multinationals

N. BEHRING/PANOS

seeking a piece of the action in China. Gan says that the company is now talking to a different US firm about a partnership that would allow ENN to deploy new technologies in the field. If it works, both companies could profit from subsequent growth and exports back to the United States and Europe.

Advocates argue that underground gasification could be one of the wisest ways to use coal, in part because it eliminates the cost — and energy — of mining and transportation. Cooking the coal in place also leaves unwanted pollutants in the ground, and any CO_2 stripped out during the chemical processing can be pumped right back where it came from. ENN isn't interested in burying CO_2 (at least not until there is money in it), although Gan is trying to diversify his energy portfolio and thus envisages using the CO_2 to stimulate the growth of algae for biofuel.

And that's the problem: unless the carbon is actually captured at its source and sequestered in some form, even the newest and fanciest coal-based liquid fuels put roughly double the CO_2 into the atmosphere compared with fuels derived from oil. China is under pressure to avoid doing exactly that, and the state-owned Shenhua Group is considering carbon capture and storage for the US$1.5-billion coal-to-liquids plant it expects to start up this year in Inner Mongolia. Shenhua is using its own technology to convert some 3.5 million tonnes of coal into diesel and other transportation fuels, equivalent to more than 24,000 barrels of oil per day. The plant will also recycle water and waste products, making it cleaner than older coal-to-liquids technologies, says Julio Friedmann, a researcher at Lawrence Livermore National Laboratory in Berkeley, California. "It's an engineering marvel."

China views the Mongolia plant as a technology showcase, and many think that Shenhua will eventually move forward with a plan to bury as much as 85% of the plant's CO_2 emissions. Without making promises, Ren Xiangkun, Shenhua's vice-president and head of its Coal Liquefaction Research Centre, says that the company attaches "great importance" to carbon management. Coal-to-liquids projects will move forward in "close connection" with the development of carbon-capture and sequestration technologies, he says. Even if CO_2 is captured during production, however, the carbon in the fuels remains. That means the best hope is to come out neutral on greenhouse emissions.

Deep burial

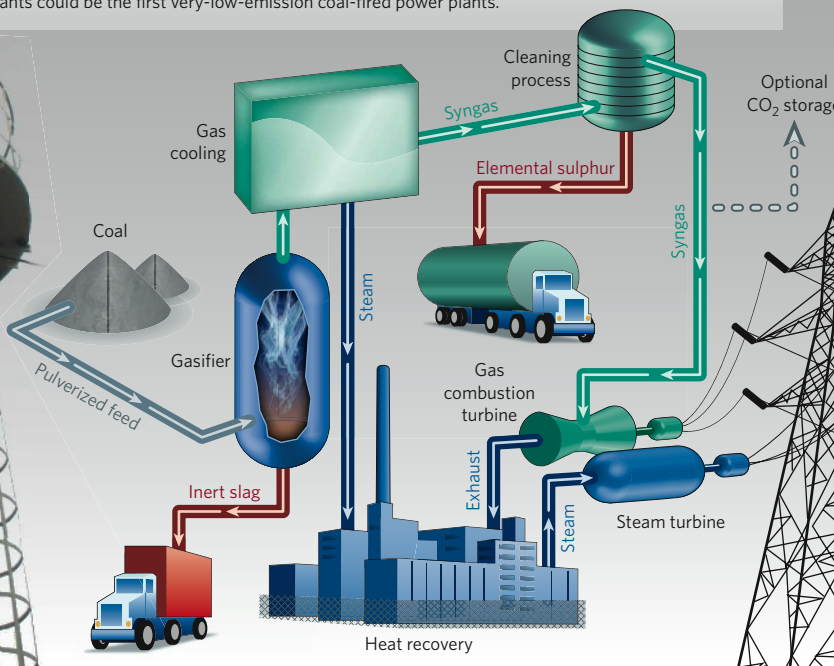
The economics seem to be enough to support sequestration. Qingyun Sun of West Virginia University in Morgantown is working with Shenhua and the US Department of Energy on the project, and says that the plant will make money as long as the price of oil is above $45 per barrel. Capturing and storing carbon emissions adds another $5–8 per barrel, but with the oil price hovering at around $130 per barrel, "that is still very profitable", Sun says. Nearby oil and gas fields could hold some of the extracted CO_2, but with volumes exceeding 3 million tonnes of CO_2 annually — larger than any sequestration project in the world so far — the ultimate target will have to be saline aquifers or deep coal seams.

Any lessons learned here might need to be applied throughout the industry. Shenhua's plant isn't even on line yet, and the company is already planning an expansion. Shenhua is also

J. TOLLEFSON

TACKLING EMISSIONS

While China works to retrofit conventional coal plants with carbon-capture technology (at left, the Gaobeidian plant in Beijing), others are working to design a next-generation coal-fired plant that sequesters most of its CO_2 emissions. Below is an outline of how an 'integrated gasification combined cycle' (IGCC) works. With the optional step of capturing and sequestering the CO_2 given off, IGCC plants could be the first very-low-emission coal-fired power plants.



Kicking the coal habit

Wei Fengrui lifts the metal lid off of an old coal-fired water boiler in a shed attached to his house. Inside are jagged lumps of coal. Two years ago he spent the equivalent of US$430 on several tonnes of coal to heat his home in Erhe Zhuang, an hour southwest of Beijing. Last winter he was able to cut his bill in half while boosting the average inside temperature from 12 to 17 °C. The trick? Insulation and double-glazed windows.

"I'm very happy," says Wei, a farmer. "It saves energy, and the rooms are warmer."

Wei's home was one of the first ten to be retrofitted under a new project led by researchers at Tsinghua University in Beijing. The team hopes to have the entire 200-residence village outfitted by next year. Insulation alone could cut coal consumption in half. If the researchers can get a facility for manufacturing pellets of biomass fuel off the ground, the village might well be able to kick the coal habit altogether, conceivably making it China's first village to wean itself off coal.

Saving a tonne or two of coal here and there might seem like a trivial pursuit, but China's rural areas are home to several hundred million people who collectively burn some 190 million tonnes of coal each year. This is equal to less than 8% of the nation's coal consumption, but only here would that be a small number. Just five countries in the world, including China itself, consume more than this.

Tsinghua professor Yang Xudong hopes the project could serve as a national model under a new government programme intended to "build a new socialist countryside". He says that the village — relatively wealthy by rural Chinese standards — is picking up 80% of the costs, which typically range from $2,400 to $3,000. Families can recoup their investment within five or six years, and so far the villagers have been receptive to the idea. "Coal is expensive, and they want to save money," Yang says. "Some families spend one-third of their income on coal."

Many Chinese villages already rely on crop residues and other biomass for energy, and Yang thinks that they have proved their ability to adopt new and varied energy solutions. Inside the same shed as Wei's boiler is his shower; a hose runs through a hole to the solar water heater on the roof. The television in his living room is a sure sign of modernity, but his bed doubles as a stove; in the winter he builds a small fire in a chamber underneath his mattress.

Switching to a new generation of high-efficiency biomass stoves would also reduce indoor air pollution, which is responsible for an estimated 380,000 deaths a year, says Kirk Smith, an expert in rural energy at the University of California, Berkeley. In some places, he says, the entire village population has been disfigured by the use of what he calls "poisonous coal" — that contaminated with fluorine and arsenic. "The simplest kind of coal stove you see, I don't think it's changed since Genghis Khan," Smith says. "This is the time to solve the problem. We've got the technology, we've got the know-how and we've got the money." **J.T.**

J. TOLLEFSON

Coal blocks are used as fuel.

partnering with South African coal-to-liquids giant Sasol to build another pair of plants that could each produce 80,000 barrels of fuel, or 3.4 million tonnes, per day. In all, seven coal-to-liquids plants are under construction in China, according to Sun, and many more are in the planning stages.

China has also been given an opportunity — one that it didn't ask for — to lead the world in developing the first low-emission coal-fuelled power plant, by coupling coal-gasification technology with carbon capture and storage. Integrated gasification combined cycle (IGCC) is a leading technology at present because the gasification process strips out conventional pollutants and produces a clean gas to generate power (see graphic). The two-stage electrical generation converts more energy into electricity, and the plant can be configured to produce a relatively pure stream of CO_2 that can be siphoned off — for a price.

Until earlier this year, the United States had been the assumed leader in the race to build the first IGCC plant with carbon capture. But in January, the US Department of Energy cancelled the signature project, called FutureGen, citing disputes with its industry partners over the $1.8-billion cost. The decision baffled and angered Chinese officials and scientists at the Huaneng Group, who were partners in the project. "This will not happen in China," says Lu Xuedu, who handles global environmental affairs as deputy chief of China's Ministry of Science and Technology. "When the Chinese government says it is going to do something, it will do it, surely."

Race for the future

With FutureGen off the table (at least in its original design or until the White House has a new occupant), the race is on between China and Australia to build the first plants. In China, Huaneng is leading a consortium that hopes to complete a 250-megawatt pilot IGCC plant by next year, then commission by 2015 a 400-megawatt plant complete with hydrogen production, fuel-cell electricity generation and carbon sequestration. Total cost: $1.5 billion, almost entirely funded by industry, although project officials say that figure could rise. The final permit has yet to be approved by China's National Development and Reform Commission (NDRC), but 'GreenGen', as it is known, has already — unofficially — broken ground along the coast in Tianjin, south of Beijing.

"China has an extensive base of real-world gasification experience."
— John Thompson

The Australians are taking a different approach with the Aus$1.2-billion (US$1.17 billion) 'ZeroGen' IGCC plant in Queensland. Project managers aim to commission a 115-megawatt pilot plant with 75% CO_2 capture and storage by 2012, followed by a 400-megawatt unit with 90% CO_2 capture by 2017. ZeroGen has already brokered agreements with local landowners and begun drilling test wells into a saline aquifer. "If we can crack it, then that has the greatest commercial application all around the world," says Chai McConnell, corporate affairs manager for the ZeroGen consortium.

Aside from GreenGen, the Chinese Ministry of Science and Technology has supported several demonstration projects that target either IGCC or 'polygeneration', which uses gasification technology to produce both power and chemicals. In addition, companies have submitted at least a dozen other IGCC applications to the NDRC, according to multiple industry sources. All these projects are pending, generating endless speculation, but few within the coal industry expect the government to approve them all. Lu thinks that as few as one or two will make it through. That's not nearly enough to make much of a difference in terms of overall greenhouse-gas emissions, given that China could bring hundreds of coal-fired plants on line in the coming years.

The problem is the cost. An analysis by Gallagher and her colleagues suggests that IGCC capital technology costs upwards of 50% more than pulverized coal in China — and that's without adding carbon capture and storage. Advanced coal power plants thus need more

government subsidies or higher electricity costs, which in turn eat into government priorities such as poverty relief and economic growth. Lu says that many companies are ready to take the lead on IGCC technology, but the government has to make its own decision. If companies fail, he says, "they come to the government saying, 'give me the money'".

Irrespective of how these initial IGCC plants fare, China will continue to develop its expertise in gasification technologies for producing chemicals. In some cases, these plants might even deploy small-scale power production on the side, a trend that some experts think is already under way. A new gasifier patented by East China University of Science and Technology, for instance, has been licensed at nearly 30 plants in China, according to the Clean Air Task Force, a US-based environmental group. "What China has going forward is an extensive base of real-world gasification experience," says John Thompson, who handles coal issues for the group. "That counts for a lot."

Even as new coal-fired plants come on line, the old ones might eventually need to be shut down or retrofitted with carbon-capture equipment. China and Australia are collaborating on several retrofit projects, including the Gaobeidian plant and some in Australia. Plenty of other pilot-scale projects are under way around the world, but the United Kingdom might well be the first country to implement post-combustion capture on a large scale. Last November, the UK government launched a competition to demonstrate 90% capture and storage on a 300–400-megawatt plant by 2014. This is also the only candidate to meet the criteria for a broader call by the European Commission for upwards of a dozen commercial-scale coal power plants with carbon capture and storage by 2015 (it's not yet clear, though, where the money will come from).

"This is what I like about GreenGen. The Chinese government decided 'we will do this' and it will be done. We in Europe rather dither about it," says Derek Taylor, who works on energy issues at the European Commission. "We've had projects announced in Europe, but none of them is that far down the line. Politicians are quiet when it comes to spending massive amounts of money on it, but massive amounts of money need to be spent."

Carbon credit

In the end, the issues faced by all these technologies are who pays, and how much. One possible source of money is the Clean Development Mechanism (CDM) of the Kyoto Protocol, which allows companies in developed nations to pay for projects that reduce emissions in developing countries. But the sums of money currently changing hands are too small. Contracts under the European emissions-trading scheme could bring in roughly $7 billion in credits to China between 2008 and 2012. That might be a hefty figure but it's not enough to affect broader energy trends in China, says Fu Ping, who works on the CDM programme under the Chinese finance ministry.

"Politicians are quiet when it comes to spending massive amounts of money."
— Derek Taylor

Moreover, the CDM programme would need to be revised and ramped up if it is to work for even one integrated carbon-storage project. First of all, it cannot currently be used to promote underground carbon storage, simply because the rules and regulations are not in place. Fu says that the United Nations, which administers the programme, has already had talks about changing that, but it might not make much of a difference until the prices for carbon storage and credits converge. CO_2 credits currently sell for $13–20 per tonne on the primary market in China, he says, and estimates for the cost of storing CO_2 from coal-fired power plants generally hovers around $50 per tonne.

Last year, the Bush administration proposed an international fund for direct investment into these types of technologies. The idea has garnered international support among the G8 leading industrialized nations, which have so far committed roughly $6 billion. Such a fund could be used to support clean-energy projects and, in the case of China, help pay for the additional costs of managing CO_2 emissions. The World Bank would manage the fund, but no decisions have been made on exactly how much money there would be or how it would be administered. Harvard's Gallagher has been advocating this approach for some time, arguing that a simpler and more aggressive tool than the CDM is needed to change the development pattern in China and other developing countries.

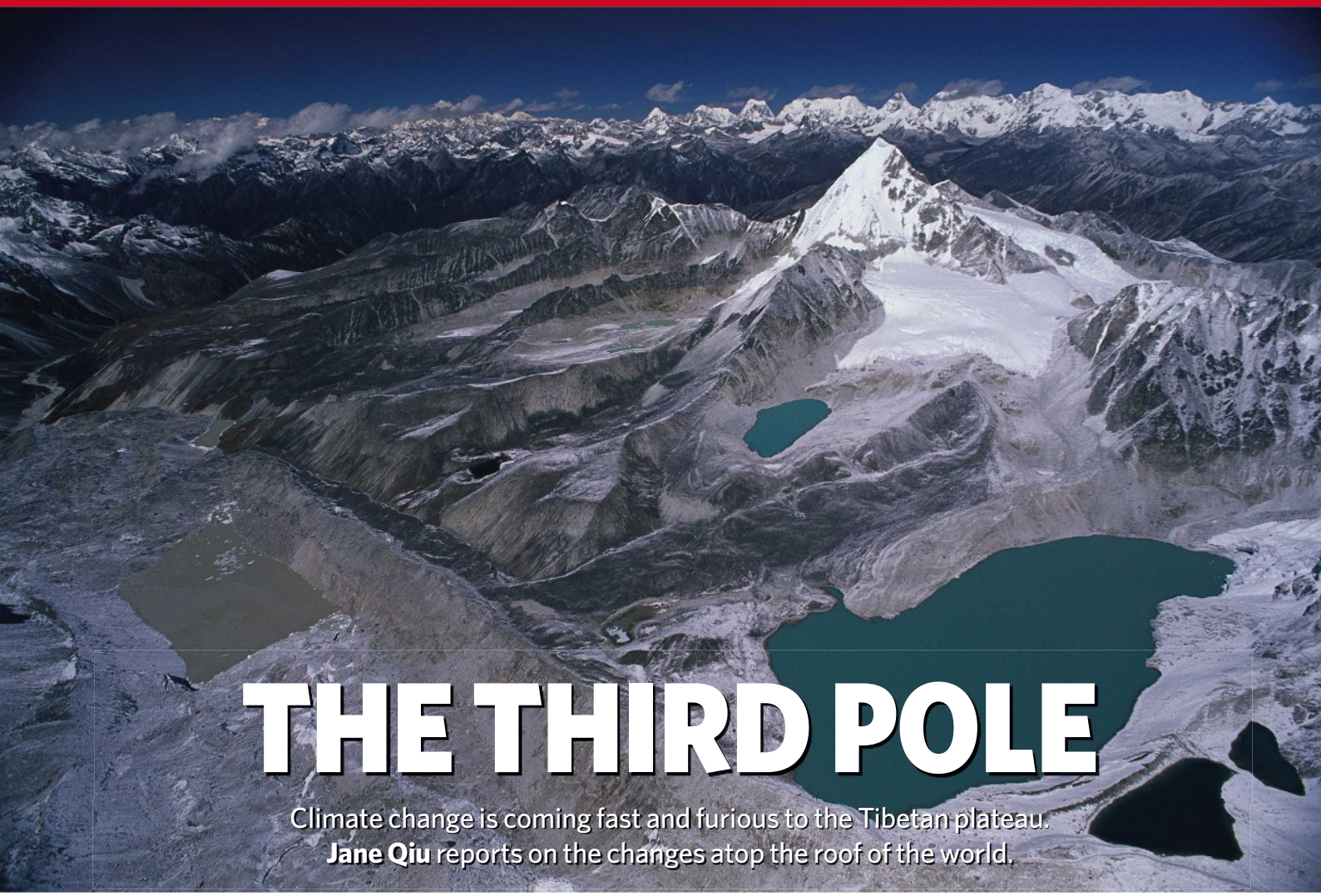
One idea being aired in climate circles in China is that the country could halve its energy intensity by 2020, then commit to levelling out emissions in the subsequent decade. A similar idea arose last year in a study sponsored by WWF China and headed by Lu and other government officials. That document also pegged the necessary investment in clean-energy technologies at roughly $220 billion. Lu plays down that number, pointing out that everything that everybody, including the Chinese, thought they knew about energy development in China five years ago was wrong.

He is also confident that China can and will tackle the problem one way or another. But if the goal is a rapid transition to a green economy, he says, the West would be wise to open the money spigot a little wider and send along its best technologies as well. "We need help," he says. ■

Jeff Tollefson covers climate from *Nature*'s Washington DC office. To hear him talk about coal in China, download the 24 July podcast at www.nature.com/podcast

See Editorial, page 367.

O. BALILTY/AP PHOTO



M. EVERTON/CORBIS

THE THIRD POLE

Climate change is coming fast and furious to the Tibetan plateau. **Jane Qiu** reports on the changes atop the roof of the world.

The Tibetan plateau gets a lot less attention than the Arctic or Antarctic, but after them it is Earth's largest store of ice. And the store is melting fast. In the past half-century, 82% of the plateau's glaciers have retreated. In the past decade, 10% of its permafrost has degraded. As the changes continue, or even accelerate, their effects will resonate far beyond the isolated plateau, changing the water supply for billions of people and altering the atmospheric circulation over half the planet.

The plateau's pivotal role is due almost entirely to its height. Being an average of 4 kilometres above sea level makes it peculiarly cold for its latitude — colder than anywhere else outside the polar regions. Lhasa, capital of the Tibet Autonomous Region, is by Tibetan standards relatively low-lying, at 3,650 metres — yet it is higher even than La Paz, Bolivia, the highest capital city of a country. Lhasa's year-round average temperature is 8 °C; at the same latitude Houston, Texas, has an average temperature of 21°C. The altitude makes Tibet cold, especially in winter; its snow and ice cover, by reflecting sunlight, make it colder still. The very bulk of the plateau affects how winds circulate above it, and its altitude also places the surface simply closer to the stratosphere than is normal.

The proximate cause of the changes now being felt on the plateau is a rise in temperature of up to 0.3 °C a decade that has been going on for fifty years — approximately three times the global warming rate. The questions are how much more change to expect in the future, and how severe the effects will be on the planet's climate as a whole. "Our understanding of global climate change would be incomplete without taking into consideration what's happening to the Tibetan plateau," says Veerabhadran Ramanathan, an atmospheric scientist at the Scripps Institution of Oceanography in La Jolla, California.

> **"A large-scale thaw of permafrost would result in the loss of its water content and trigger an ecological catastrophe."**
> — Ouyang Hua

Perhaps surprisingly given its significance, the potential impact of the Tibetan plateau is still unfamiliar to many climatologists. One reason is that there are far fewer data available compared with the Arctic and Antarctic, which have seen a far greater number of scientific expeditions to plumb their secrets. Although fieldwork there can be tough, the plateau offers the same physical isolation coupled with political challenges, at least for Western researchers. "The plateau's remoteness, high altitude and harsh weather conditions make any research on the region very challenging," says Yao Tandong, director of the Institute of Tibetan Plateau Research, headquartered in Beijing, of the Chinese Academy of Sciences.

Yao and his colleagues should know: in the 1980s, they were among the few researchers persevering in difficult field conditions to gather data on the plateau's past climate history. They drilled ice cores, up to 300 metres long, from Himalayan glaciers 7,200 metres high. "It's all done manually, and we had to carry them down the mountain. There were no helicopters, no heavy equipment," he says. "It's -30 °C, with the wind cutting through us like a knife. It's no mean feat." Such ordeals seem to have paid off: in collaboration with glaciologist Lonnie Thompson of Ohio State University in Columbus, the team's work on oxygen isotopes within the cores yielded the most comprehensive temperature reconstruction for the plateau, showing a large-scale warming trend that began in the twentieth century and is amplified at higher elevations[1]. Their findings are consistent with temperature records from meteorological

NATURE|Vol 454|24 July 2008

Lifting the roof of the world

The rise of the Tibetan plateau is thought to have intensified the Indian monsoon. So the history of when and how it rose could improve the understanding of climate history and long-term climate change.

Some think that the plateau rose in blocks, progressively from south to north, after the India subcontinent collided into Asia some 50 million years ago. Others suspect that this neat model might be overly simplistic. To resolve this debate, researchers have studied the elevation history of various sites on the plateau.

Geologists have inferred elevation from the composition of oxygen isotopes in ancient rain and snow that fell on the plateau, and are preserved in rocks and lake sediments. The method is based on the observation that the higher a mountain range is, the less oxygen-18 is precipitated, whereas the opposite holds for oxygen-16. Thus, the $^{18}O/^{16}O$ ratio can be used to deduce past elevation. Studies of sediments at various sites on the plateau show that those areas were at elevations of over 4,000 metres between 11 million and 35 million years ago[11–13].

Another way of estimating elevation is to look at the shape and size of fossil leaves (pictured, right). "The leaves of a plant represent an engineering solution to a set of environmental constraints, which is dictated by laws of physics and chemistry," says Robert Spicer, a geologist at the Open University in Milton Keynes, UK. Spicer and his colleagues found that atmospheric enthalpy — a measure of energy that depends on both temperature and moisture — was recorded in fossil leaves and could be used as a direct readout of the elevation at which the plant grew. Studies of fossilized leaves from more than 20 species from the Namling basin, in the southern Tibetan plateau, show that the elevation of the region 15 million years ago was more than 4,600 metres[14].

More recently, Wu Zhenhan, a geologist at the Chinese Academy of Sciences's Institute of Geomechanics in Beijing, and his colleagues studied two groups of ancient lake basins in the central Tibetan plateau at 4,500 metres, which were excavated during the construction of the Qinghai-Tibet railway[15]. "They are beautifully horizontal, probably untouched for about 20 million years," says Wu.

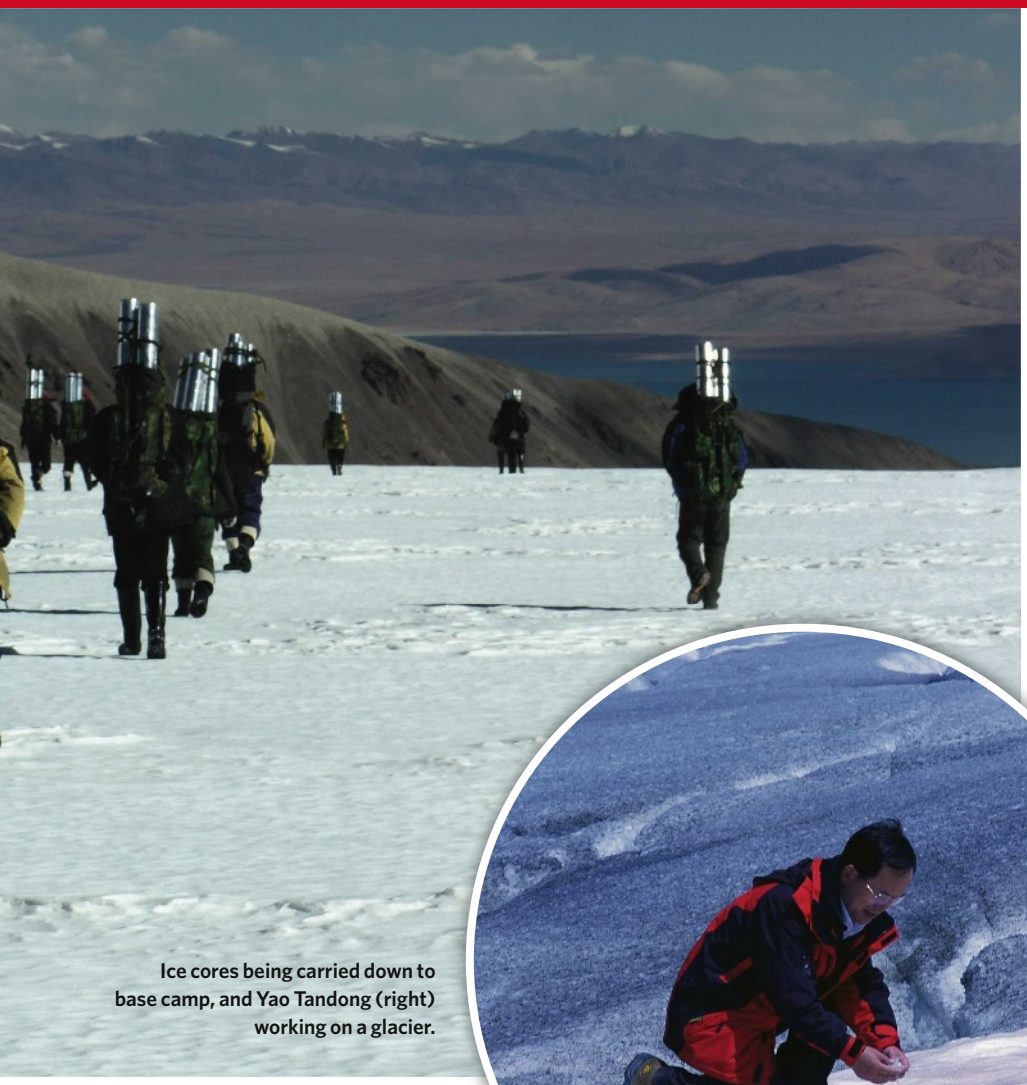
In the early Miocene time, between about 22 million and 17 million years ago, the lake basins formed two gigantic lake complexes of 100,000 and 50,000 square kilometres. "It's unlikely that lakes of that size could be uplifted to the present elevation without any distortion," he says. This suggests that the central plateau was 4,500 metres above sea level as early as 20 million years ago. **J.Q.**

R. SPICER/OPEN UNIV.

stations that have made continuous measurements since the 1950s [ref. 2].

Some of this is what you would expect in a world undergoing greenhouse warming, but there are regional factors on the plateau that exacerbate the effect. In summer, dust from regional deserts blows towards and up against the northern and southern slopes of the plateau. One recent satellite study, for instance, tracked dust wafting in from the Taklamakan desert to the north[3]. "We were really surprised to find this much dust over the plateau," says Huang Jianping, an atmospheric scientist at Lanzhou University and lead author of the study. The dust layers can reach as high as 10 kilometres above sea level, where they both absorb and reflect sunlight, changing the amount of radiation that reaches the plateau.

Combining with the dust to drive climate change are emissions of 'black carbon', the soot that results when people cook with biofuels such as wood, crop waste or dung. Southeast Asia, including the Himalayas, is one of the global hotspots for black-carbon emissions[4]. Using unmanned aircraft, Ramanathan and colleagues measured the amount of sunlight absorbed by black carbon, and found that it contributes as much as 50% of the solar heating of the air[5]. "It's the second-largest contributor to atmospheric warming in the region, after



Ice cores being carried down to base camp, and Yao Tandong (right) working on a glacier.

carbon dioxide," he says. He estimates that the combined effect of black carbon and greenhouse gases may be sufficient to account for a warming trend of 0.25 °C per decade in the Himalayas, roughly what has been observed so far.

When black carbon settles on Himalayan glaciers, it darkens the snow and ice so that they absorb more heat and become warmer. "The melting seasons on the plateau now begin earlier and last longer," says Xu Baiqing of the Institute of Tibetan Plateau Research. Glaciers at the edge of the plateau tend to melt more than those in the middle; one study, for instance, showed that glaciers in the eastern part of the Kunlun Mountains retreated by 17% over the past 30 years, which is ten times faster than those in the central plateau. If current trends hold, two-thirds of the plateau glaciers could be gone by 2050, says Yao.

Floods and droughts

The melting glaciers are starting to leave behind dangerous glacial lakes, in which meltwater ponds behind a dam of debris left by the retreating ice tongue. Scientists have identified 34 such glacial lakes on the northern slopes of the Himalayas, and 20 outburst floods have been recorded in the past 50 years.

The risk of floods, though, is but a short-term danger far exceeded by long-term issues with water supplies atop the plateau. Runoff from the region's mountains feeds the largest rivers across Southeast Asia, including the Yangtze, Yellow, Mekong, Ganges and Indus rivers. If glaciers continue to retreat and snowpack shrinks atop the plateau, the water supplies of billions of people will be in danger[6].

Permafrost is also at risk, as rising temperatures cause the 'active' ground layer — which freezes and thaws every year — to thicken. That, in turn, affects how heat and moisture flow between the ground and the atmosphere, further perturbing the system[7]. Degradation of permafrost will not only put the Qinghai–Tibet Railway at risk[8], but also endangers the plateau's alpine ecosystems, which rely on permafrost to trap water in the topmost layers of soil to allow plants to thrive at an altitude that would otherwise be too hostile for them. "A large-scale thaw of permafrost would result in the loss of its water content and trigger an ecological catastrophe," says Ouyang Hua, deputy director of the Institute of Geographical Sciences and Natural Resources Research in Beijing. As permafrost stores one-third of the world's soil carbon, vegetation loss would lead to a huge amount of carbon entering the atmosphere, exacerbating global warming.

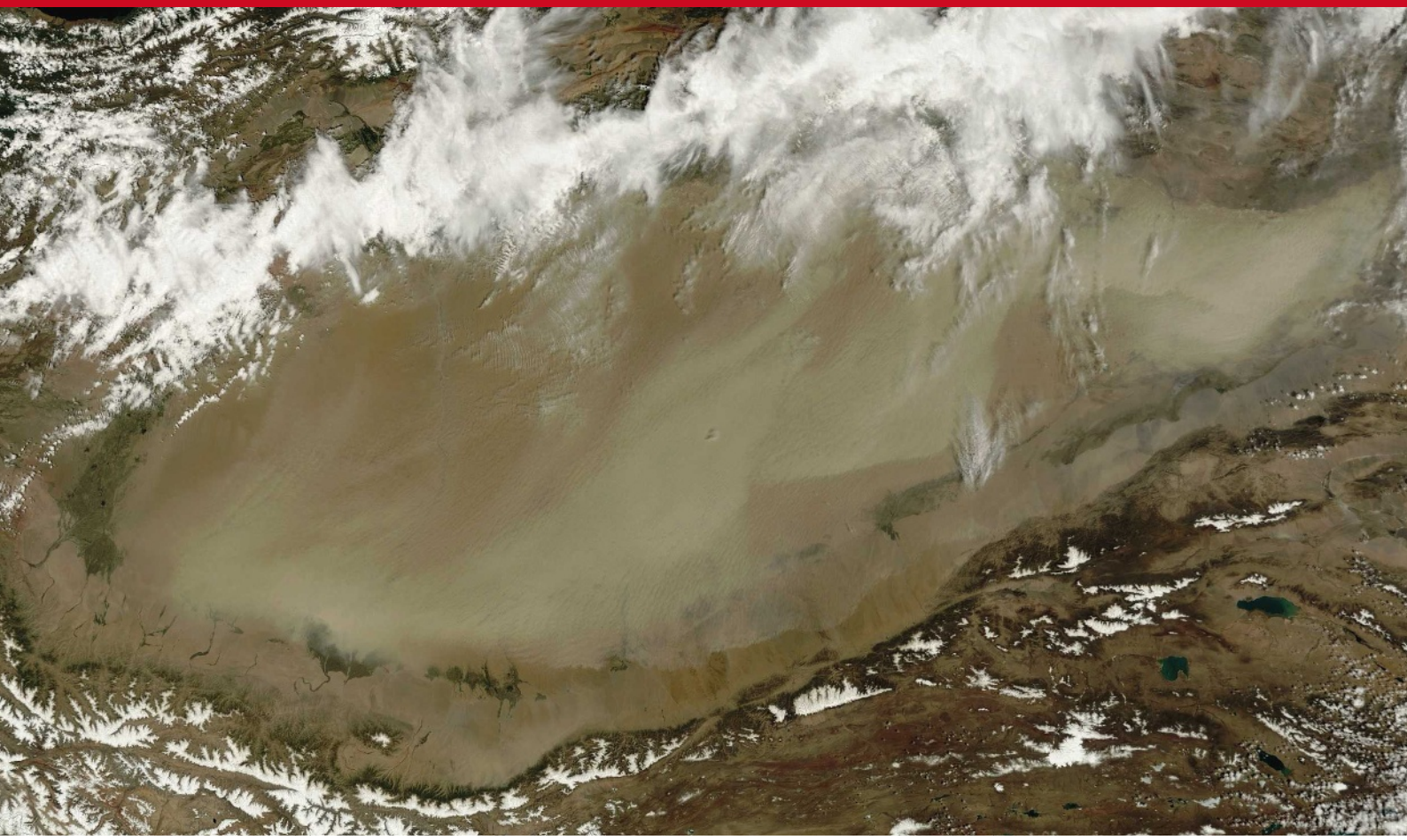
Competing forces

With all the changes the Tibetan plateau is undergoing — a warming climate, retreating glaciers, degrading permafrost and alpine ecosystems — what are the implications for the regional and global climate? The first and most important victim could be the Indian monsoon. This strong seasonal wind results from differences in the thermal properties between land and ocean. In summer, the vast land in Asia heats up more than the Indian Ocean, leading to a pressure gradient and the flow of the air and moisture from the ocean. The rise of the Tibetan plateau starting 50 million years ago (see 'Lifting the roof of the world') is thought to have strengthened this effect. As the land surface absorbs more sunlight than the atmosphere, the plateau creates a vast area of surface warmer than the air at that elevation, thereby increasing the land–ocean pressure gradient and intensifying the monsoon.

Some climate models show that global warming would lead to a greater increase in the plateau's surface temperature than over the ocean, thus augmenting the monsoon. On the other hand, some models suggest that aerosols that absorb solar radiation, and changes in land use in the region, could weaken the monsoon. "The intensity of the monsoon is likely to depend on which of these two competing forces dominates," says Ramanathan.

No matter what the causes are, some studies indicate that the weakening force may be prevailing, or has prevailed for at least the past three centuries. Duan Keqin, of the Cold and Arid Regions Environmental and Engineering Research Institute in Lanzhou, and his colleagues reconstructed a 300-year history of snow accumulation by analysing ice cores from the Dasuopu glacier[9]. They believe the ice there preserves an estimate of monsoon variations in the Himalayas. "We found that the warmer it was, the weaker the monsoon," says Duan. On average, a temperature increase of 0.1 °C was associated with a decrease of 100 millimetres in snow accumulation. But similar studies on other parts of the plateau are needed to confirm the results, he notes.

"Changes in the Indian monsoon are not the



Plumes from dust storms in the Taklamakan desert, such as this one in June 2005, can reach the Tibetan plateau and affect the climate there.

NASA/MODIS RAPID RESPONSE TEAM

only threat in Asia to the global climate," adds Rong Fu of the Georgia Institute of Technology in Atlanta. Her research shows that convection over the Tibetan plateau can transport water vapour and pollutants to the stratosphere[10], the atmospheric layer that is immediately above the troposphere and contains most of the Earth's ozone. "The strong, horizontal wind in the stratosphere could then spread the water vapour and pollutants globally," says Fu.

Water vapour has a stronger greenhouse effect than carbon dioxide per molecule, but it normally reaches no higher than 1–2 kilometres below the stratosphere. The situation is different over the plateau, over which the convection layer is shifted some 6 kilometres further up so that its top boundary is around 18 kilometres up, in the lower stratosphere. In addition, the troposphere is thinner over the plateau, and the heat emitted by the surface can reach higher and make the air warmer at the base of the stratosphere. "So more water vapour is able to get to the stratosphere without being frozen or precipitated," says Fu. Warmer temperatures over the plateau can resulting increased glacial melting and water-vapour transport — which, in turn, causes strong convection and lifts even more water vapour up. "It's very worrying to think that a lot of it may reach the stratosphere," she says.

"Reducing emissions of greenhouse gases and black carbon should be the top priority."
— Xu Baiqing

"Worrying", indeed, best captures the mood of researchers who work on the Tibetan plateau. They are keen to undertake large-scale, comprehensive studies and to collect as many data as possible. "We know so little about it and understand it even less," says Yao. One ongoing study is to document all the glaciers in China, recording characteristics such as their location, area, length, thickness and the position of the snow line. A similar survey was conducted between 1978 and 2002, which scientists believe could serve as a reference point to reveal any major changes. In addition, glaciologists continue to identify and closely monitor potentially dangerous glacial lakes in hopes of heading off any potential outburst floods.

Quick way out

Meanwhile, others focus on the bigger picture of how to tackle pollution problems in Asia. "Reducing emissions of greenhouse gases and black carbon should be the top priority," says Xu. Ramanathan reckons that cutting down on black-carbon emissions could be a "quick way out of the mess", given that its half-life in the atmosphere is about 15–20 days compared with the century-scale half-life of carbon dioxide. His simulations suggest that, just by removing traditional ways of cooking with wood, dung and crop residues, some 40–60% of the black-carbon emissions would be gone. This could be "a short-term fix, a low-hanging fruit that is much cheaper and faster" than reducing carbon dioxide, he says. "The key is to give villagers access to better forms of energy."

In the end, the Tibetan plateau may be a crucial testing ground for how humans and the environment collide in a globally warmed world. Can the world's third pole be saved? "Let's hope that the changes the plateau is going through are only transient," says Yao. "What we do about them probably will determine what's going to happen to it in the future." ■

Jane Qiu writes for *Nature* from Beijing.

1. Yao, T. *et al. Annals Glaciol.* **43,** 1–7 (2006).
2. Liu, X. & Chen, B. *Int. J. Climatol.* **20,** 1729–1742 (2000).
3. Huang, J. *et al. Geophys. Res. Lett.* **34,** L18805 (2007).
4. Ramanathan, V. & Carmichael, G. *Nature Geosci.* **1,** 221–227 (2008).
5. Ramanathan, V. *et al. Nature* **448,** 575–578 (2007).
6. Cyranoski, D. *Nature* **438,** 275–276 (2005).
7. Cheng, G. & Wu, T. *J. Geophy. Res.* **112,** F02S03 (2007).
8. Qiu, J. *Nature* **449,** 398–402 (2007).
9. Duan, K., Yao, T. & Thompson, L. G. *J. Geophys. Res.* **111,** D19110 (2006).
10. Fu, R. *et al. Proc. Natl. Acad. Sci. USA* **103,** 5664–5669 (2006).
11. Garzione, C. N., Dettman, D. L., Quade, J., DeCelles, P. G. & Butler, R. F. *Geology* **28,** 339–342 (2000).
12. Rowley, D. B. & Currie, B. S. *Nature* **439,** 677–681 (2006).
13. DeCelles, P. G. *et al. Earth and Planet. Sci. Lett.* **253,** 389–401 (2007).
14. Spicer, R.A. *et al. Nature* **421,** 622–624 (2003).
15. Wu, Z. *et al. Geol. Soc. Am. Bull.* doi: 10.1130/B26043.1 (2008).

See Editorial, page 367, and News Feature, page 384.

CORRESPONDENCE

China's move to higher-meat diet hits water security

SIR — Your Editorial 'A fresh approach to water' (*Nature* **452,** 253; 2008) points out that the world's looming water crisis is driven by climate change, population growth and economic development. In China, changing food-consumption patterns are the main cause of the worsening water scarcity. If other developing countries follow China's trend towards protein-rich Western diets, the global water shortage will become still more severe.

In China, it takes 2,400–12,600 litres of water to produce a kilogram of meat, whereas a kilogram of cereal needs only 800–1,300 litres (J. Liu and H. H. G. Savenije *Hydrol. Earth Syst. Sci.* **12,** 887–898; 2008). The recent rise in meat consumption has pushed China's annual per capita water requirement for food production up by a factor of 3.4 from 255 cubic metres in 1961 to 860 cubic metres in 2003. Compared with China's population growth by a factor of 1.9 over the same period, this suggests that dietary change is making a high demand on water resources.

China's water requirement for food production is still well below that of many developed countries. The United States, for example, uses 1,820 cubic metres per capita per year. But the steady increase in the amount of meat in Chinese diets is worrying. Consumption already exceeds by 50% the optimal amount recommended by the Chinese Nutrition Society — although discrepancies between rural and urban areas and between eastern and western regions are significant. This diet shift may also have detrimental effects on the population's health, as in developed countries. In general, changes in food-consumption patterns are closely related to affluence, although they are influenced by food preferences as well. Raising public awareness about healthy eating habits could also help to mitigate water scarcity.

Junguo Liu and Hong Yang Swiss Federal Institute of Aquatic Science and Technology, Ueberlandstrasse 133, PO Box 611, CH-8600, Dübendorf, Switzerland
e-mail: water21water@yahoo.com
H. H. G. Savenije Delft University of Technology, Department of Water Management, PO Box 5048, 2601 DA, Delft, The Netherlands,
See Editorial, page 367

In the wake of two retractions, a request for investigation

SIR — Your Editorial 'Negative results' (*Nature* **453,** 258; 2008) and News Feature 'Designer debacle' (*Nature* **453,** 275–278; 2008), on the retraction of two papers from my laboratory and the events surrounding those retractions, have provided opportunity for misunderstanding and misinterpretation.

Regrettably, as with all human endeavours, mistakes can occur in scientific research. When a mistake is made, it should be admitted through retraction of the published paper. Such retractions should then lead to a process of impartial scientific enquiry and analysis, as well as introspection by the participants.

I have acknowledged, and will continue to acknowledge, my personal responsibility to the scientific community for these errors as well as my responsibility as a research supervisor to my students.

As my actions have been called into question, I have asked the Duke University Medical Center administration to hold a formal and impartial inquiry into these retractions and the events that have followed. My request has been granted by the university.

Homme W. Hellinga Duke University Medical Center, Durham, North Carolina 27710, USA
e-mail: hwh@biochem.duke.edu

Fusion needs a realistic cost assessment

SIR — In your Editorial about the increasing expense of the ITER fusion reactor ('The price isn't right' *Nature* **453,** 824; 2008), you suggest that scientists might be suspected of deliberately under-quoting the price to help sell the project. Possibly. What then should one make of the projected costs of fusion energy outlined in the European Fusion Development Agency in its 2005 report (see http://tinyurl.com/5gvh5o)?

The report gives a projected electricity cost for a 1.5-gigawatt plant of conservative design of €0.09 (US$0.14) per kilowatt-hour. This is rather high compared with, say, renewables; but it goes on to state (on the basis of untested conceptual designs with less conventional materials) that this cost would be reduced to €0.05 per kilowatt-hour "in a mature fusion industry". This figure is only a little higher than for conventional nuclear power plants. Moreover, it has been quoted by leaders in the fusion community (see C. Llewellyn-Smith and D. Ward, *Nuclear Future* **2,** 93–100; 2006). Hidden in these projected costs is that both the €0.09 and €0.05 numbers have already been reduced by a factor of 0.65 to give the cost of 'a tenth of a kind' — that is, for the tenth reactor. The original cost estimate of nearly €0.14 per kilowatt-hour has been cleverly reduced by almost a factor of three. At a time when the economics of fusion energy needed some support, one cannot help admiring this approach.

As the costs of ITER come under further pressure on all sides and the huge technical problems become recognized, there is an urgent need for realistic and independent assessment of the costs of a practical fusion device, and even as to whether it is sensible for the programme to continue. Apart from the plasma conditions, what about the tritium breeding and reprocessing, the massive materials problems in a radiation-damage environment, the near-impossibility of maintenance and the difficulty of maintaining the integrity of superconducting magnets over long periods? The list of areas where heroic engineering is needed goes on and on.

As US physicist William Metz once said, "It sometimes seems necessary to suspend one's normal critical faculties not to find the problems of fusion overwhelming." The fusion story is like a snowball going downhill gathering mass and momentum — impossible to stop, and at the end there will be only a pool of water to show for all the effort.

J. H. Evans Abingdon, Oxfordshire, UK
e-mail: jhevans@sky.com

Fewer academics are not the answer to funding woes

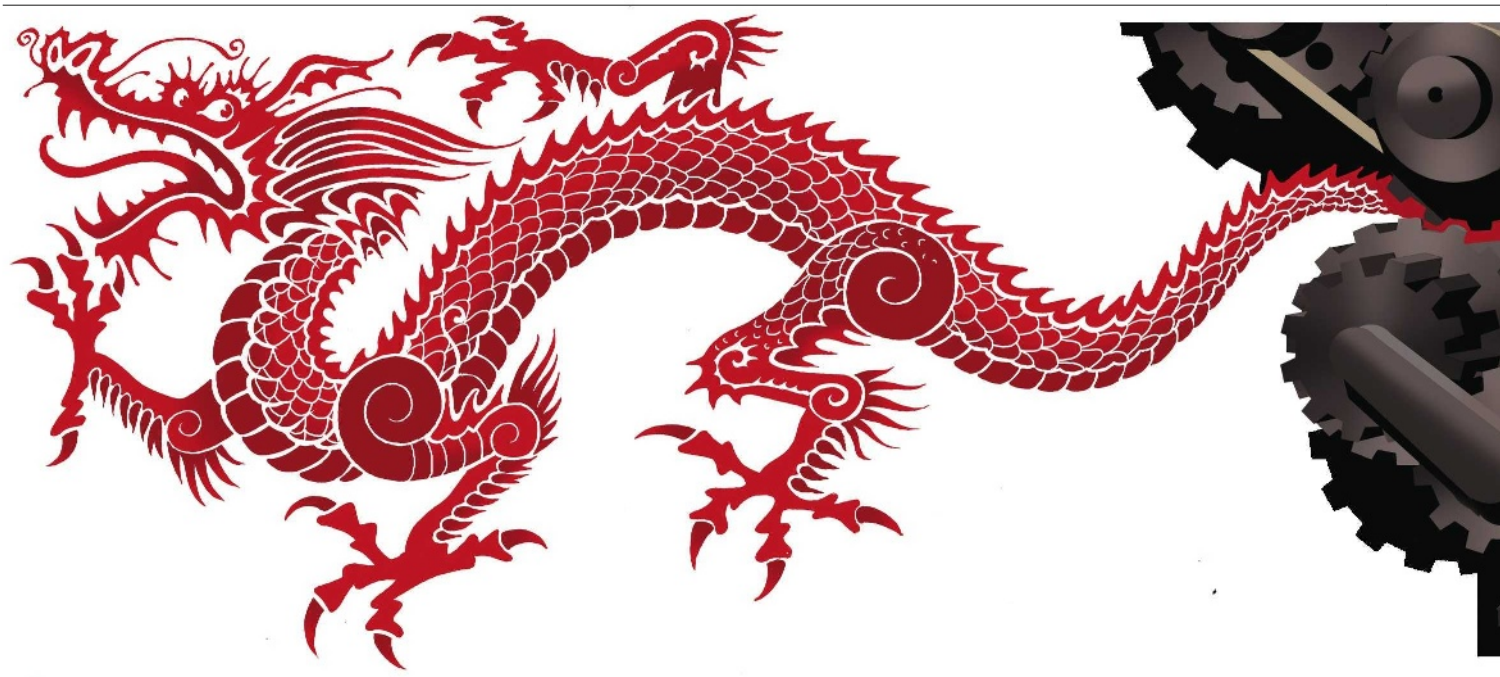
SIR — In his Correspondence 'Fewer academics could be the answer to insufficient grants' (*Nature* **453,** 978; 2008), Andrew Doig suggests that the endemic problem of the rejection of high-quality grant proposals could be solved by cutting the number of academic staff. This proposal could create a new problem.

The number of academic staff is generally related to the number of undergraduates. Cutting the number of academics would reduce the number of trained students produced, which would have a negative effect on the nation's health and wealth.

In this increasingly technological age, we need all the trained scientists we can muster to combat issues such as global warming. The way to prevent the rejection of high-quality grant proposals and to support research is to put a bit more money into the system.

Philip Strange School of Pharmacy, University of Reading, Reading, Berkshire RG6 6AJ, UK
e-mail: p.g.strange@reading.ac.uk

COMMENTARY



The prizes and pitfalls of progress

Pushes to globalize science must not threaten local innovations in developing countries argues **Lan Xue**.

Developing countries such as China and India have emerged both as significant players in the production of high-tech products, and as important contributors to the production of ideas and global knowledge. China's rapid ascent as a broker rather than simply a consumer of ideas and innovation has made those in the 'developed' world anxious. A 2007 report by UK think tank Demos says that "US and European pre-eminence in science-based innovation cannot be taken for granted. The centre of gravity for innovation is starting to shift from west to east"[1].

But the rapid increase in research and development spending in China — of the order of 20% per year since 1999 — does not guarantee a place as an innovation leader. Participation in global science in developing countries such as China is certainly good news for the global scientific community. It offers new opportunities for collaboration, fresh perspectives and a new market for ideas. It also presents serious challenges for the management of innovation in those countries. A major discovery in the lab does not guarantee a star product in the market. And for a country in development, the application of knowledge in productive activities and the related social transformations are probably more important than the production of the knowledge itself. By gumming the works in information dissemination, by misplacing priorities, and by disavowing research that, although valuable, doesn't fit the tenets of modern Western science, developing countries may falter in their efforts to become innovation leaders.

Vicious circle

China's scientific publications (measured by articles recorded in the Web of Science) in 1994 were around 10,000, accounting for a little more than 1% of the world total. By 2006, the publications from China rose to more than 70,000, increasing sevenfold in 12 years and accounting for almost 6% of the world total (see graph, page 400). In certain technical areas, the growth has been more dramatic. China has been among the leading countries in nanotechnology research, for example, producing a volume of publications second only to that of the United States.

The publish-or-perish mentality that has arisen in China, with its focus on Western journals, has unintended implications that threaten to obviate the roughly 8,000 national scientific journals published in Chinese. Scientists in developing countries such as China and India pride themselves on publishing articles in journals listed in the Science Citation Index (SCI) and the Social Science Citation Index (SSCI) lists. In some top-tier research institutions in China, SCI journals have become the required outlet for research.

A biologist who recently returned to China from the United States was told by her colleague at the research institute in the prestigious Chinese Academy of Sciences (CAS) that publications in Chinese journals don't really count toward tenure or promotion. Moreover, the institute values only those SCI journals with high impact factors. Unfortunately, the overwhelming majority of the journals in SCI and SSCI lists are published in developed countries in English or other European languages. The language requirement and the high costs of these journals mean that few researchers in China will have regular access to the content. Thus as China spends more and publishes more, the results will become harder to find for Chinese users. This trend could have a devastating impact on the local scientific publications and hurt China's ability to apply newly developed knowledge in an economically useful way.

Several members of the CAS expressed their concerns on this issue recently at the 14th CAS conference in Beijing. According to Molin Ge,

ILLUSTRATIONS BY D. PARKINS

a theoretical physicist at the Chern Institute of Mathematics, Nankai University, Tianjin, as more high-quality submissions are sent to overseas journals, the quality of submissions to local Chinese journals declines, which lowers the impact of the local Chinese journals. This becomes a vicious circle because the lower the impact, the less likely these local journals are to get high-quality submissions[2].

Setting agendas

Research priorities in developing countries may be very different from those in developed nations, but as science becomes more globalized, so too do priorities. At the national level, developing countries' research priorities increasingly resemble those of the developed nations, partly as a result of international competitive pressures. For example, after the United States announced its National Nanotechnology Initiative (NNI) in 2001, Japan and nations in Europe followed suit, as did South Korea, China, India and Singapore. According to a 2004 report by the European Union[3], public investment in nanotechnology had increased from €400 million (US$630 million) in 1997 to more than €3 billion in 2004.

In their words

Researchers and businesspeople in China, expatriates and 'returnees' give their views of what it will take to make China a research and innovation powerhouse.

CHINA

Ling-An Wu
Professor, Institute of Physics, Chinese Academy of Sciences, Beijing

Fix the gender ratio

When I returned to China in 1962 I was impressed by the equality of men and women in society. Even during the 'Cultural Revolution' there was no prejudice against women, although political discrimination was routine. The reforms of the 1980s opened a new era for science, yet contrary to expectation female scientists have not fared so well. In physics the situation is particularly discouraging. Formerly, 25% of the research staff at our institute were female, but that has dropped to 14%, while the percentage of women full professors has fallen from 17% to 7%. In the physics department of Tsinghua University in Beijing, the percentage of retired female full professors is 19% whereas that of those currently employed is 8%.

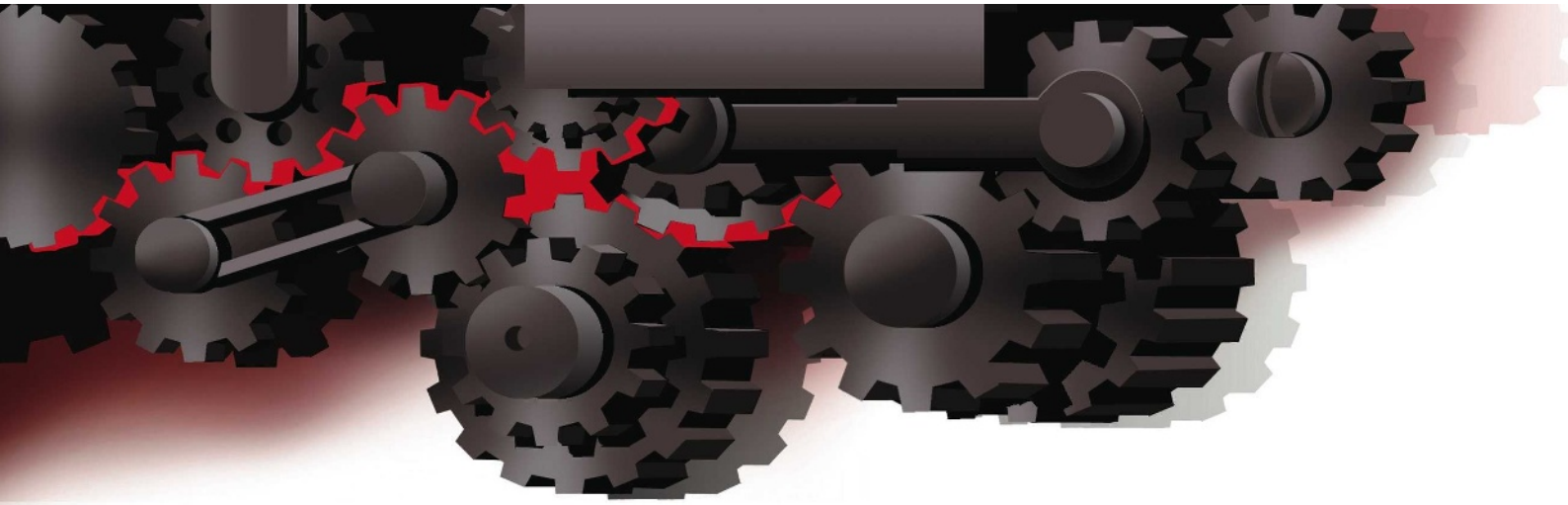
Discrimination now menaces both younger and older women: some employers openly declare that only male applicants need apply, while many institutions force women of associate professor status to retire at age 55; their male counterparts can retire at 60. The current (predominantly male) leadership is not concerned with the statistics. It is true that the number of female postgraduate students has risen, but the chief reason is that job discrimination everywhere is pushing them to seek higher degrees. Will this new generation be able to find their way to the top in China, or will they pursue better opportunities abroad, or just be wasted along the leaky pipeline?

Wolfgang Hennig
CAS-MPG Partner Institute for Computational Biology, Shanghai, China and Johannes Gutenberg-University Mainz, Germany

Overhaul education

My first experience of China was in 1985 and since then I've taught in the Chinese Academy of Sciences in Shanghai and elsewhere on behalf of the Max Planck Society, the German Academic Exchange Service and the Chinese Academy of Sciences to improve biological sciences training for Chinese students and to create contacts with European students. So, it's a good time to ask what has changed in China in two decades. The obvious answer is everything — from the thousands of bikes now replaced by air-choking cars to the boost in funding and the focus on science and technology.

Still, what has not changed during the past 20 years is the educational approach in China. It is based on memorizing and reproducing knowledge rather than on developing one's own initiative, critical thinking and originality. Postdocs trained in China rarely show the ability to work independently or demonstrate creativity in the selection of and approach to research subjects. While I had excellent Chinese students in the past, educated in my lab in the Netherlands, today many of the highly qualified students move into commercial fields through business schools and management courses. Making money has become the major attraction in China and this has severe consequences at the university level: basic research is not considered as important and attractive as it had been. Considering the living conditions of most students — dormitories still house four to six students to a room without heating or air conditioning — one can understand this desire.

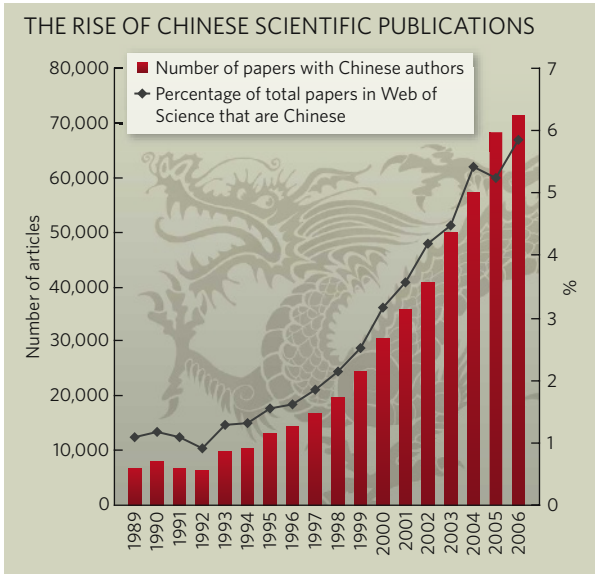


Part of the pressure to jump on the international bandwagon comes from researchers themselves. Scientists in the developing world maintain communications with those elsewhere. It is only natural that they want to share the attention that their colleagues in the developed Western world and Japan are receiving by pursuing the same hot topics. The research is exciting, fast-moving and often easier to publish. At the same time, there are many other crucial challenges to be met in developing countries. For example, public health, water and food security, and environmental protection all beg for attention and resources. If people perceive these research areas as less intellectually challenging and rewarding, the issues will fail to receive the resources, support and recognition they require. Without better agenda-setting practices, the scientific community will continue to face stinging criticism. It can send a satellite to Mars but not solve the most basic problems that threaten millions of lives in the developing world.

The introduction of Western scientific ideals to the developing world can generate an environment that is hostile to the indigenous research that prima facie does not fit those ideals. The confrontation between Western medicine and traditional Chinese medicine dates back to the early days of the twentieth century when Western medicine was first introduced in China. The debate reached a peak last year when a famous actress, Xiaoxu Chen, died from breast cancer. She allegedly insisted on treatment by Chinese traditional medicine, raising the hackles of some who claimed it to be worthless. Many Chinese still support traditional medicine and say that the dominance of Western medicine risks endangering China's scientific and cultural legacy.

A similar row erupted around earthquake prediction. In the 1960s and 1970s, China set up a network of popular earthquake-prediction stations, using simple instruments and local knowledge. For the most part, the network was decommissioned as China built the modern earthquake-monitoring system run by the China Earthquake Administration. When the system failed to predict the recent Sichuan earthquake, several people claimed that non-mainstream approaches had predicted its imminence. Scientists in the agency have tended to brush off such unofficial and individual predictions. To many this seems arrogant and bureaucratic.

It would be foolish and impossible to stop the globalization of science. There are tremendous benefits to science enterprises in different countries being integrated into a global whole. One should never think of turning back the clock. At the same time, it is possible to take some practical steps to minimize the harmful effects of this trend on local innovation.



Prioritizing for the people

First of all, there is a need to re-examine the governance of global science in recognition of the changing international geography of science. Many international norms and standards should be more open and accommodating to the changing environment in developing countries. For example, there is a need to re-evaluate the SCI and SSCI list of journals to include quality journals in the developing countries. In the long run, the relevant scientific community could also think about establishing an international panel to make decisions on the selection of journals for these indices, given their important influence. The recent move by Thomson Reuters, the parent company of ISI, to expand its coverage of the SCI list by adding 700 regional academic journals, is a step in the right direction[4].

English has become the de facto global language of science. Developing countries should invest in public institutions to provide translation services so that global scientific progress can be disseminated quickly. Developing countries can learn from Japan, a world leader in collecting scientific information and making it available to the public in the local language. At the same time, there should also be international institutions to provide similar services to the global

science community so that "results and the knowledge generated through research should be freely accessible to all", as advocated by Nobel Laureates John Sulston and Joseph Stiglitz[5].

When setting agendas, governments in developing countries must be careful in allocating their resources for science to achieve a balance between following the science frontier globally and addressing crucial domestic needs. A balance should also be struck between generating knowledge and disseminating and using knowledge. In addition, the global science community has a responsibility to help those developing countries that do not have adequate resources to solve problems themselves.

Finally, special efforts should be made to differentiate between pseudoscience and genuine scientific research. For the latter, one should tolerate or even encourage such indigenous research efforts in developing countries even if they do not fit the recognized international science paradigm. After all, the real advantage of a globalized scienctific enterprise is not just doing the same research at a global scale, but doing new and exciting research in an enriched fashion. ■

Lan Xue is in the School of Public Policy and Management, and the director of the China Institute for Science and Technology Policy, Tsinghua University, Beijing 100084, China.
e-mail: xuelan@tsinghua.edu.cn

1. Leadbeater, C. & Wilsdon, J. *The Atlas of Ideas: How Asian Innovation Can Benefit Us All* (Demos, 2007).
2. Xie, Y. *et al. Good submissions went overseas — Chinese S&T journals could not keep up with their overseas peers* Chinese Youth Daily, 25 June 2008.
3. http://ec.europa.eu/nanotechnology/pdf/nano_com_en_new.pdf
4. http://scientific.thomsonreuters.com/press/2008/8455931/
5. Sulston, J. & Stiglitz, J. Science is being held back by outdated laws, *The Times* (5 July 2008).

See Editorial, page 367, and News Feature, page 382.
To comment on this article and others in our innovation series, visit http://tinyurl.com/5uolx2.

Li Gong

Chief executive, Mozilla Online, Beijing

Liberate funding

Research funding in China has increased manyfold in the past two decades. The National Natural Science Foundation awards alone went from 80 million renminbi (US$12 million) in 1986 to 4.33 billion renminbi in 2007. These programmes have succeeded in nurturing researchers and yielding research papers, books and patents. However, regulations governing how research funds can be spent, established in part to prevent misuse and abuse, are handicapping researchers and institutions, distorting research activities, and resulting in significant waste.

Chinese research programmes operate much like those in the rest of the industrial world, with a major difference that China has strict national guidelines on project spending. The most flawed rule dictates that researchers are already paid salaries and thus only a small portion of funding (usually 10–20%, sometimes less) can be spent on personnel. In reality, academic research salaries are uncompetitive. The average overall income per faculty member at top computer science departments is comparable to a fresh graduate's starting salary at IBM or Microsoft. These government guidelines make it much more profitable to stay outside the academic institutions, and drive researchers towards more commercial projects to earn more income. The rules can also result in spending on equipment that is unnecessary and in the worst cases resold, and in wasteful conferences and trips, meals and entertaining, and other excesses. The spending regulations are a significant drag on research performance, fund efficiency, and personal advancement, and need urgent reform.

Cong Cao

Sociologist, Neil D. Levin Graduate Institute of International Relations and Commerce, State University of New York

Encourage returnees

Of the some 1.2 million Chinese who have gone abroad as students and scholars, only a quarter have returned, thereby constituting an unequivocal 'brain drain' for China. Indeed, non-returnees, especially academics, are most likely to be the best and the brightest, who are most needed in China's innovation push.

Besides taking several years to set up a laboratory, form a team, recruit students, apply and get grants, and start the research, returned academics have to adapt and adjust to a 'different' research environment and be involved in various activities unimaginable to those abroad. They risk not being able to survive because they do not know the rules of the game played in China, and without *guanxi* — personalized networks of influence — and social and political connections, they have no one to turn to for help.

The costs of working in China are high. Some productive scientists have expressed the wish to return permanently and demonstrate that it is possible to do first-rate science in China. But this depends on whether China can provide the kind of research environment that will help them thrive.

Jianguo Liu

Director, Center for Systems Integration and Sustainability, Michigan State University, East Lansing; guest professor, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences (CAS)

Integrate disciplines

China's unprecedented economic boom and societal changes have created unexpected environmental challenges. Divorce, for example, usually splitting one household into two smaller ones, is increasingly common in China and traditional multi-generation families are also fragmenting. More households require more land and construction material for housing. Smaller households are often inefficient and produce relatively more wastes and pollutants. To tackle environmental problems, there is an urgent need to integrate natural sciences with socioeconomics, demography, human behavior and policy, addressing seemingly unrelated trends. Enhancing international partnerships for systems integration is a win-win strategy for China and other nations.

a theoretical physicist at the Chern Institute of Mathematics, Nankai University, Tianjin, as more high-quality submissions are sent to overseas journals, the quality of submissions to local Chinese journals declines, which lowers the impact of the local Chinese journals. This becomes a vicious circle because the lower the impact, the less likely these local journals are to get high-quality submissions[2].

Setting agendas

Research priorities in developing countries may be very different from those in developed nations, but as science becomes more globalized, so too do priorities. At the national level, developing countries' research priorities increasingly resemble those of the developed nations, partly as a result of international competitive pressures. For example, after the United States announced its National Nanotechnology Initiative (NNI) in 2001, Japan and nations in Europe followed suit, as did South Korea, China, India and Singapore. According to a 2004 report by the European Union[3], public investment in nanotechnology had increased from €400 million (US$630 million) in 1997 to more than €3 billion in 2004.

In their words

Researchers and businesspeople in China, expatriates and 'returnees' give their views of what it will take to make China a research and innovation powerhouse.

CHINA

Ling-An Wu
Professor, Institute of Physics, Chinese Academy of Sciences, Beijing

Fix the gender ratio

When I returned to China in 1962 I was impressed by the equality of men and women in society. Even during the 'Cultural Revolution' there was no prejudice against women, although political discrimination was routine. The reforms of the 1980s opened a new era for science, yet contrary to expectation female scientists have not fared so well. In physics the situation is particularly discouraging. Formerly, 25% of the research staff at our institute were female, but that has dropped to 14%, while the percentage of women full professors has fallen from 17% to 7%. In the physics department of Tsinghua University in Beijing, the percentage of retired female full professors is 19% whereas that of those currently employed is 8%.

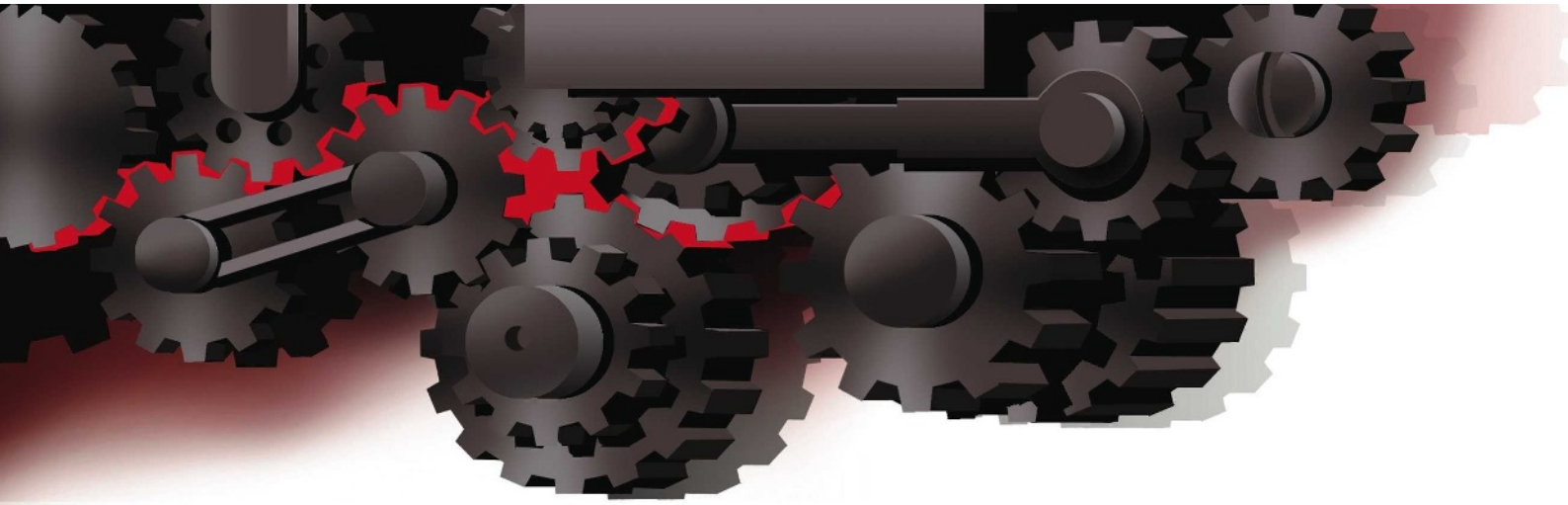
Discrimination now menaces both younger and older women: some employers openly declare that only male applicants need apply, while many institutions force women of associate professor status to retire at age 55; their male counterparts can retire at 60. The current (predominantly male) leadership is not concerned with the statistics. It is true that the number of female postgraduate students has risen, but the chief reason is that job discrimination everywhere is pushing them to seek higher degrees. Will this new generation be able to find their way to the top in China, or will they pursue better opportunities abroad, or just be wasted along the leaky pipeline?

Wolfgang Hennig
CAS-MPG Partner Institute for Computational Biology, Shanghai, China and Johannes Gutenberg-University Mainz, Germany

Overhaul education

My first experience of China was in 1985 and since then I've taught in the Chinese Academy of Sciences in Shanghai and elsewhere on behalf of the Max Planck Society, the German Academic Exchange Service and the Chinese Academy of Sciences to improve biological sciences training for Chinese students and to create contacts with European students. So, it's a good time to ask what has changed in China in two decades. The obvious answer is everything — from the thousands of bikes now replaced by air-choking cars to the boost in funding and the focus on science and technology.

Still, what has not changed during the past 20 years is the educational approach in China. It is based on memorizing and reproducing knowledge rather than on developing one's own initiative, critical thinking and originality. Postdocs trained in China rarely show the ability to work independently or demonstrate creativity in the selection of and approach to research subjects. While I had excellent Chinese students in the past, educated in my lab in the Netherlands, today many of the highly qualified students move into commercial fields through business schools and management courses. Making money has become the major attraction in China and this has severe consequences at the university level: basic research is not considered as important and attractive as it had been. Considering the living conditions of most students — dormitories still house four to six students to a room without heating or air conditioning — one can understand this desire.

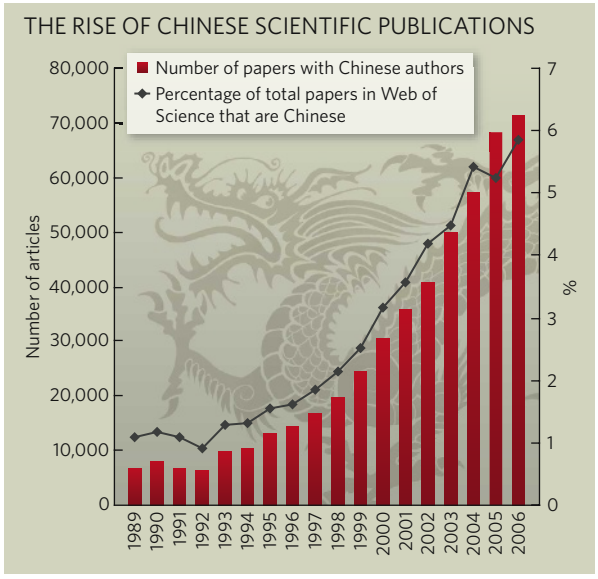


Part of the pressure to jump on the international bandwagon comes from researchers themselves. Scientists in the developing world maintain communications with those elsewhere. It is only natural that they want to share the attention that their colleagues in the developed Western world and Japan are receiving by pursuing the same hot topics. The research is exciting, fast-moving and often easier to publish. At the same time, there are many other crucial challenges to be met in developing countries. For example, public health, water and food security, and environmental protection all beg for attention and resources. If people perceive these research areas as less intellectually challenging and rewarding, the issues will fail to receive the resources, support and recognition they require. Without better agenda-setting practices, the scientific community will continue to face stinging criticism. It can send a satellite to Mars but not solve the most basic problems that threaten millions of lives in the developing world.

The introduction of Western scientific ideals to the developing world can generate an environment that is hostile to the indigenous research that prima facie does not fit those ideals. The confrontation between Western medicine and traditional Chinese medicine dates back to the early days of the twentieth century when Western medicine was first introduced in China. The debate reached a peak last year when a famous actress, Xiaoxu Chen, died from breast cancer. She allegedly insisted on treatment by Chinese traditional medicine, raising the hackles of some who claimed it to be worthless. Many Chinese still support traditional medicine and say that the dominance of Western medicine risks endangering China's scientific and cultural legacy.

A similar row erupted around earthquake prediction. In the 1960s and 1970s, China set up a network of popular earthquake-prediction stations, using simple instruments and local knowledge. For the most part, the network was decommissioned as China built the modern earthquake-monitoring system run by the China Earthquake Administration. When the system failed to predict the recent Sichuan earthquake, several people claimed that non-mainstream approaches had predicted its imminence. Scientists in the agency have tended to brush off such unofficial and individual predictions. To many this seems arrogant and bureaucratic.

It would be foolish and impossible to stop the globalization of science. There are tremendous benefits to science enterprises in different countries being integrated into a global whole. One should never think of turning back the clock. At the same time, it is possible to take some practical steps to minimize the harmful effects of this trend on local innovation.



Prioritizing for the people

First of all, there is a need to re-examine the governance of global science in recognition of the changing international geography of science. Many international norms and standards should be more open and accommodating to the changing environment in developing countries. For example, there is a need to re-evaluate the SCI and SSCI list of journals to include quality journals in the developing countries. In the long run, the relevant scientific community could also think about establishing an international panel to make decisions on the selection of journals for these indices, given their important influence. The recent move by Thomson Reuters, the parent company of ISI, to expand its coverage of the SCI list by adding 700 regional academic journals, is a step in the right direction[4].

English has become the de facto global language of science. Developing countries should invest in public institutions to provide translation services so that global scientific progress can be disseminated quickly. Developing countries can learn from Japan, a world leader in collecting scientific information and making it available to the public in the local language. At the same time, there should also be international institutions to provide similar services to the global

science community so that "results and the knowledge generated through research should be freely accessible to all", as advocated by Nobel Laureates John Sulston and Joseph Stiglitz[5].

When setting agendas, governments in developing countries must be careful in allocating their resources for science to achieve a balance between following the science frontier globally and addressing crucial domestic needs. A balance should also be struck between generating knowledge and disseminating and using knowledge. In addition, the global science community has a responsibility to help those developing countries that do not have adequate resources to solve problems themselves.

Finally, special efforts should be made to differentiate between pseudoscience and genuine scientific research. For the latter, one should tolerate or even encourage such indigenous research efforts in developing countries even if they do not fit the recognized international science paradigm. After all, the real advantage of a globalized scienctific enterprise is not just doing the same research at a global scale, but doing new and exciting research in an enriched fashion. ■

Lan Xue is in the School of Public Policy and Management, and the director of the China Institute for Science and Technology Policy, Tsinghua University, Beijing 100084, China.
e-mail: xuelan@tsinghua.edu.cn

1. Leadbeater, C. & Wilsdon, J. *The Atlas of Ideas: How Asian Innovation Can Benefit Us All* (Demos, 2007).
2. Xie, Y. *et al. Good submissions went overseas — Chinese S&T journals could not keep up with their overseas peers* Chinese Youth Daily, 25 June 2008.
3. http://ec.europa.eu/nanotechnology/pdf/nano_com_en_new.pdf
4. http://scientific.thomsonreuters.com/press/2008/8455931/
5. Sulston, J. & Stiglitz, J. Science is being held back by outdated laws, *The Times* (5 July 2008).

See Editorial, page 367, and News Feature, page 382.
To comment on this article and others in our innovation series, visit http://tinyurl.com/5uolx2.

Li Gong

Chief executive, Mozilla Online, Beijing

Liberate funding

Research funding in China has increased manyfold in the past two decades. The National Natural Science Foundation awards alone went from 80 million renminbi (US$12 million) in 1986 to 4.33 billion renminbi in 2007. These programmes have succeeded in nurturing researchers and yielding research papers, books and patents. However, regulations governing how research funds can be spent, established in part to prevent misuse and abuse, are handicapping researchers and institutions, distorting research activities, and resulting in significant waste.

Chinese research programmes operate much like those in the rest of the industrial world, with a major difference that China has strict national guidelines on project spending. The most flawed rule dictates that researchers are already paid salaries and thus only a small portion of funding (usually 10–20%, sometimes less) can be spent on personnel. In reality, academic research salaries are uncompetitive. The average overall income per faculty member at top computer science departments is comparable to a fresh graduate's starting salary at IBM or Microsoft. These government guidelines make it much more profitable to stay outside the academic institutions, and drive researchers towards more commercial projects to earn more income. The rules can also result in spending on equipment that is unnecessary and in the worst cases resold, and in wasteful conferences and trips, meals and entertaining, and other excesses. The spending regulations are a significant drag on research performance, fund efficiency, and personal advancement, and need urgent reform.

Cong Cao

Sociologist, Neil D. Levin Graduate Institute of International Relations and Commerce, State University of New York

Encourage returnees

Of the some 1.2 million Chinese who have gone abroad as students and scholars, only a quarter have returned, thereby constituting an unequivocal 'brain drain' for China. Indeed, non-returnees, especially academics, are most likely to be the best and the brightest, who are most needed in China's innovation push.

Besides taking several years to set up a laboratory, form a team, recruit students, apply and get grants, and start the research, returned academics have to adapt and adjust to a 'different' research environment and be involved in various activities unimaginable to those abroad. They risk not being able to survive because they do not know the rules of the game played in China, and without *guanxi* — personalized networks of influence — and social and political connections, they have no one to turn to for help.

The costs of working in China are high. Some productive scientists have expressed the wish to return permanently and demonstrate that it is possible to do first-rate science in China. But this depends on whether China can provide the kind of research environment that will help them thrive.

Jianguo Liu

Director, Center for Systems Integration and Sustainability, Michigan State University, East Lansing; guest professor, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences (CAS)

Integrate disciplines

China's unprecedented economic boom and societal changes have created unexpected environmental challenges. Divorce, for example, usually splitting one household into two smaller ones, is increasingly common in China and traditional multi-generation families are also fragmenting. More households require more land and construction material for housing. Smaller households are often inefficient and produce relatively more wastes and pollutants. To tackle environmental problems, there is an urgent need to integrate natural sciences with socioeconomics, demography, human behavior and policy, addressing seemingly unrelated trends. Enhancing international partnerships for systems integration is a win-win strategy for China and other nations.

Ming-Wei Wang
Director, the National Center for Drug Screening, Shanghai, Professor of Pharmacology, Shanghai Institute of Materia Medica, CAS

Build bridges
China's drive to transform its pharmaceutical industry from imitation to innovation has reaped its first fruit: consistent elevation of research standards, rapid growth of a talent pool, and massive build-up of essential technology platforms.

Apart from common technical difficulties and social hurdles shared with Western peers, Chinese pharmaceutical scientists are facing some unique challenges. First, discovery activities are largely conducted at academic institutions. Chinese drug companies are much less interested in early lead compounds than their Western counterparts. Second, the current regulatory system is not optimized for innovation — it is far more difficult and takes much longer to file an investigational new drug application in China than in the United States. Third, public tolerance for failure is not a major part of the culture. High expectation and pressure to produce short-term returns often force scientists to abandon long-term and more difficult projects. Working as a fast follower or contract service provider is thus a convenient strategy.

International collaboration, especially with reputable industrial partners, could shorten the time for learning and reduce trial errors thereby strengthening China's drug-discovery capabilities. With more transnational pharmaceutical giants setting up R&D mechanisms in China, innovation-oriented cooperations are expected to increase.

Duanqing Pei
Professor and deputy director, Guangzhou Institute of Biomedicine and Health, CAS

Arm the troops better
Researchers are like an army fighting a tough war with imported, somewhat unreliable, weapons. Most reagents and equipment must be shipped from abroad, and thus are only available in leading research centres in major cities. Perishable supplies are handled by careless local suppliers. Domestic companies have sprouted up, but they need time to learn how to serve the biosciences sector.

Xiang Yu
Professor, Institute of Neuroscience, CAS

Be patient
As a young scientist returning to China after nearly 20 years of studying and working abroad, the task of building a research group is daunting and exciting. The usual challenges of starting to run a laboratory without prior experience are compounded by attempting the task in a country where the education and research systems are themselves undergoing a merger between East and West, traditional and modern.

Many students were attracted to biology during college because it was a popular choice and are choosing graduate school because of the tight job market. Breaking habits is difficult as these students are the first generation from 'one child families', each growing up with two parents and four grandparents attending only to them. Why should they bother listening to me, the only 'parent' in the lab, airing Western ideas and making things up as I go along? Well, it's an experiment. It is exciting to witness history in the making and be part of it. As to how my experiment is going, please check back in 20 years.

Zhangjie Shi
Associate professor, College of Chemistry, Peking University, Beijing

Foster collaboration
Historically it has been hard for researchers to establish interdisciplinary collaborations. Pure synthetic chemistry has grown fast, but fields such as chemical biology face major challenges because of gaps between chemical and biological research. Few scientists in China can really bridge chemistry and biology, in contrast to those in the United States. Fortunately, many people are working to change this. The rapid pace of reform and change makes me believe that these problems will be solved or partially solved by efforts from government and scientific communities in China in the future.

Liping Wei
Professor, Center of Bioinformatics, College of Life Sciences, Peking University

Tolerate risks
In this time of fast economic growth, it is easy to discount innovations in basic sciences that do not necessarily have clear and immediate practical value. Funding agencies must tolerate higher risks and longer research cycles.

"A closed society is less likely to produce internationally recognized scientific achievements, and this is particularly true for a developing country such as China that had long been isolated from the rest of the world."
— Xing Xu, palaeontologist, Institute of Vertebrate Paleontology and Paleonanthropology, Chinese Academy of Sciences

BOOKS & ARTS

How one child was deemed enough

Scientific policy-making in China has come a long way since the 1970s, argue **Ling Chen** and **Gang Zhang**.

Just One Child: Science and Policy in Deng's China
by Susan Greenhalgh
University of California Press: 2008.
426 pp. $55.00, £32.95 (hbk)
$21.95, £12.95 (pbk)

CHINA

Beijing's subway trains are always packed with locals, migrant workers, job seekers and tourists. Most are in their twenties or thirties; they look well fed, well clothed, healthy, educated and wealthy. Some read newspapers, others talk on mobile phones. Having recently read *Just One Child*, one looks at these passengers with new eyes, trying to imagine the scene if China had not adopted its 'one child' policy. Would the trains have been more crowded, the people less wealthy? Would they have been happier or wiser if they had brothers or sisters?

China's family-planning policy, in place for almost three decades, ranks with its economic reforms as among the most transformative measures in modern Chinese history. The policy, rolled out in 1980, deemed that each couple could have just one offspring or face penalties.

People from Western cultures might assume that the controversial policy was just one in a long line of inscrutable political movements within China's authoritarian leadership of the period. According to anthropologist Susan Greenhalgh's new book, however, the real impetus for this sweeping measure came from a small group of ambitious aeronautical scientists who ventured into population modelling in 1980. Using predictions of population growth from computer models, the group convinced China's political leadership that only drastic action would control the country's soaring number of citizens. Greenhalgh's investigation of the history and politics of this fateful policy decision is meticulous. Through compelling storytelling and analysis, she draws together field and archival studies that cover the two decades from 1982 to 2007, spanning huge social, political, cultural and geographical distances.

The family-planning policy was launched just as the spirit of science was recovering from the country's traumatic Cultural Revolution, at a time when Chinese people's enthusiasm to advance economic development and transform society had reached unprecedented heights. The booming population seemed to be a growing burden for the whole country.

PRIVATE COLLECTION/DATO IMAGES/BRIDGEMAN ART LIBRARY

Birth-control campaigns were boosted by 'proof' from population models by missile scientists.

Social scientists and demographers suggested several milder methods of population control, but it was leading scientists from the Ministry of Aeronautics and Astronautics who played a decisive role in the policy-making. These researchers had achieved great success during the 1960s and 1970s working on missile projects that used mathematical control theory, gaining them the trust of politicians. When they used their models on a social-science issue for the first time, their results were treated as facts.

Greenhalgh contends that the advice of the missile scientists, thanks to their seemingly authoritative scientific tools, fortified the conviction of Chinese leaders that strict family planning was a national necessity.

Through her penetrating analysis, the author attempts to explore relationships between the state, science, technology and society, and the rise of modern China. She argues that science was far from a tool manipulated by politicians. Rather, she contends that the scientists involved exerted their influence through normal decision-making channels, such as persuading veteran cadres and planning departments to agree with their suggestions. Extrapolating from the case of the family-planning policy, she seems to warn that the intrusion of science into sociopolitical fields can be rather dangerous.

In fact, the situation was more complex: the flaw was unscientific policy-making. In late-1970s China, the governance and process of policy-making were far from well established. Its political leaders decided that population control was crucial to the future of the country, leaving them no choice but to embrace the policy. They then sought a scientific 'proof' to help the idea gain popular acceptance, but did not fully debate the alternatives. China's political system was unable to address the bias of scientism.

Chinese policy-making remains a mystery to many natives, let alone to foreigners. Happily, the process has changed significantly during the past decades. Before Deng Xiaoping's economic reforms and opening up of China from 1978, policy-making involved only national political leaders, top technocrats and scientific elites. During the 1980s and 1990s, Deng's reforms strengthened policy-making roles and awareness in governmental and non-governmental organizations and social-interest groups.

The current government, headed by President Hu Jintao and Premier Wen Jiabao, puts even greater emphasis on policies that place the value of people and their human rights at their centre. This government also emphasizes a scientific concept of development, which refers to a rational and sustainable growth model that balances economic development, social welfare and environmental and

ecological sustainability, and aims to address the damaging social, ecological and environmental effects of the current growth model. It encourages public organizations to take part in crafting social, environmental and industrial policies. For example, the recent reforms to the health-care system involved more than six organizations, including universities, research institutions, foreign consulting companies and international bodies, who submitted proposals that were then debated within and outside the government. Science is playing an increasingly important role beyond providing justifications for government policies.

Chinese policy-makers now have dramatically different educational backgrounds and characteristics from their predecessors. Between the 1950s and 1960s, veteran soldiers with limited education held almost all major government posts. In the 1980s, Deng Xiaoping's new criteria of cadre selection promoted middle-aged officials with engineering backgrounds into senior positions, resulting in a government dominated by technocrats. In the 1990s, those with degrees and experience in economics and public management gradually moved to the centre of politics. Hopefully, future generations of Chinese policy-makers will be equipped with social, political and legal knowledge conducive to an enhanced understanding of the human impact of public policies and the significance of scientific policy-making.

The family-planning policy has had both negative and positive effects on Chinese society. It has produced an alarmingly wide gender gap in the sector of the population born after the 1980s, and an inverted pyramid demographic that will be challenging to care for in the coming decades. The effects of a generation of 'little kings' on Chinese society and culture remain to be seen. However, the policy seems to have helped China move into the fast-lane of economic development. It may also have accelerated the improvement of the population's well-being, as evinced by higher education levels and lower infant mortality rates.

In reality, the family-planning policy was never fully implemented. Ethnic minorities and rural peoples — the majority of China's population — could in practice have two or more children, if not by policy design, then by paying an economic, political and social cost, such as in lost public-sector jobs or heavy fines. And from 1984, rural residents whose first child was female were allowed to have a second child. China's real fertility was thus estimated to be around 1.8 children per family in 2006. However, according to a study in 2006, there are no accurate data because of missing birth registration records that have resulted in a hidden population.

It is now vital to determine if the policy should be relaxed, and what should succeed it. In recent years, Chinese demographers and policy-makers have begun to try to identify a fertility rate that would balance the population. This time it seems more likely that China will set a rational policy, having much improved its scientific policy-making system. ■

Ling Chen is an assistant professor in the School of Public Policy and Management, Tsinghua University, Beijing 100084, China.
e-mail: chenling@tsinghua.edu.cn
Gang Zhang is principal administrator in the Directorate for Science, Technology and Industry at the Organisation for Economic Co-operation and Development, F-75775 Paris Cedex 16, France.

See Editorial, page 367, and News Special Report, page 374.

RTKL ASSOCIATES

Beijing's new science centre will be the first of many in China.

A museum in every district

China Science and Technology Museum
Olympic Village, Beijing
Opening September 2009

CHINA

With up to 30,000 visitors a day, the Beijing-based China Science and Technology Museum is grossly oversubscribed. In response, China is building another one more than twice the size, costing 2 billion yuan (US$300 million). The museum (artist's impression, pictured) will open in Beijing's Olympic Village in September 2009, in a building designed to resemble an ancient Chinese puzzle, the Lock of Luban.

The museum will showcase scientific and technological developments in all disciplines, from agriculture, geology, alternative energy and environmental protection to space exploration, as well as inventions from ancient China. There will be an exhibition hall for children, who are expected to constitute half of the visitors. With running costs of 150 million yuan a year, the building will boast the world's largest dome video screen and laboratories where participants can do short research projects.

"The new museum is emblematic of China's long-term commitment to science communication," says Zhu Youwen, director of the venue's planning and development. In June 2002 China's top legislature, the Standing Committee of the National People's Congress, passed a bill on the dissemination of developments in science and technology to the public.

The country's current 15-year strategic plan for science and technology, announced in February 2006, prioritizes the improvement of public understanding in these areas. Infrastructure is the first step. All 34 districts in China plan to have at least one science museum in their capital cities by 2010, adding to the 40 or so already in existence. More than a dozen are under construction, including what will be the world's largest science museum when it opens in Guangzhou, Guangdong province.

Many applaud China's political and financial commitment to science communication. But some critics, such as science historian Liu Bing of the Centre of Science, Technology and Society at Beijing's Tsinghua University, are concerned that the quality of exhibitions and events may not be up to scratch. Some provincial science museums also fail to attract significant visitor numbers, and there are few public debates on topical or controversial issues such as traditional Chinese medicine, stem-cell therapies and genetically modified crops.

Zhu concedes that there is much room for improvement. So Beijing's new science museum will foster closer collaborations with its counterparts elsewhere in the country and abroad. It also plans to host seminars and workshops at which scientists, policy-makers and the public can debate crucial scientific matters.

Some hurdles must still be overcome, Zhu explains. Public participation in Chinese policy-making is a new concept to all involved. Government officials are not yet reconciled to having to justify political decisions to the populace, nor

ecological sustainability, and aims to address the damaging social, ecological and environmental effects of the current growth model. It encourages public organizations to take part in crafting social, environmental and industrial policies. For example, the recent reforms to the health-care system involved more than six organizations, including universities, research institutions, foreign consulting companies and international bodies, who submitted proposals that were then debated within and outside the government. Science is playing an increasingly important role beyond providing justifications for government policies.

Chinese policy-makers now have dramatically different educational backgrounds and characteristics from their predecessors. Between the 1950s and 1960s, veteran soldiers with limited education held almost all major government posts. In the 1980s, Deng Xiaoping's new criteria of cadre selection promoted middle-aged officials with engineering backgrounds into senior positions, resulting in a government dominated by technocrats. In the 1990s, those with degrees and experience in economics and public management gradually moved to the centre of politics. Hopefully, future generations of Chinese policy-makers will be equipped with social, political and legal knowledge conducive to an enhanced understanding of the human impact of public policies and the significance of scientific policy-making.

The family-planning policy has had both negative and positive effects on Chinese society. It has produced an alarmingly wide gender gap in the sector of the population born after the 1980s, and an inverted pyramid demographic that will be challenging to care for in the coming decades. The effects of a generation of 'little kings' on Chinese society and culture remain to be seen. However, the policy seems to have helped China move into the fast-lane of economic development. It may also have accelerated the improvement of the population's well-being, as evinced by higher education levels and lower infant mortality rates.

In reality, the family-planning policy was never fully implemented. Ethnic minorities and rural peoples — the majority of China's population — could in practice have two or more children, if not by policy design, then by paying an economic, political and social cost, such as in lost public-sector jobs or heavy fines. And from 1984, rural residents whose first child was female were allowed to have a second child. China's real fertility was thus estimated to be around 1.8 children per family in 2006. However, according to a study in 2006, there are no accurate data because of missing birth registration records that have resulted in a hidden population.

It is now vital to determine if the policy should be relaxed, and what should succeed it. In recent years, Chinese demographers and policy-makers have begun to try to identify a fertility rate that would balance the population. This time it seems more likely that China will set a rational policy, having much improved its scientific policy-making system. ■

Ling Chen is an assistant professor in the School of Public Policy and Management, Tsinghua University, Beijing 100084, China.
e-mail: chenling@tsinghua.edu.cn
Gang Zhang is principal administrator in the Directorate for Science, Technology and Industry at the Organisation for Economic Co-operation and Development, F-75775 Paris Cedex 16, France.

See Editorial, page 367, and News Special Report, page 374.

RTKL ASSOCIATES

Beijing's new science centre will be the first of many in China.

A museum in every district

China Science and Technology Museum
Olympic Village, Beijing
Opening September 2009

CHINA

With up to 30,000 visitors a day, the Beijing-based China Science and Technology Museum is grossly oversubscribed. In response, China is building another one more than twice the size, costing 2 billion yuan (US$300 million). The museum (artist's impression, pictured) will open in Beijing's Olympic Village in September 2009, in a building designed to resemble an ancient Chinese puzzle, the Lock of Luban.

The museum will showcase scientific and technological developments in all disciplines, from agriculture, geology, alternative energy and environmental protection to space exploration, as well as inventions from ancient China. There will be an exhibition hall for children, who are expected to constitute half of the visitors. With running costs of 150 million yuan a year, the building will boast the world's largest dome video screen and laboratories where participants can do short research projects.

"The new museum is emblematic of China's long-term commitment to science communication," says Zhu Youwen, director of the venue's planning and development. In June 2002 China's top legislature, the Standing Committee of the National People's Congress, passed a bill on the dissemination of developments in science and technology to the public.

The country's current 15-year strategic plan for science and technology, announced in February 2006, prioritizes the improvement of public understanding in these areas. Infrastructure is the first step. All 34 districts in China plan to have at least one science museum in their capital cities by 2010, adding to the 40 or so already in existence. More than a dozen are under construction, including what will be the world's largest science museum when it opens in Guangzhou, Guangdong province.

Many applaud China's political and financial commitment to science communication. But some critics, such as science historian Liu Bing of the Centre of Science, Technology and Society at Beijing's Tsinghua University, are concerned that the quality of exhibitions and events may not be up to scratch. Some provincial science museums also fail to attract significant visitor numbers, and there are few public debates on topical or controversial issues such as traditional Chinese medicine, stem-cell therapies and genetically modified crops.

Zhu concedes that there is much room for improvement. So Beijing's new science museum will foster closer collaborations with its counterparts elsewhere in the country and abroad. It also plans to host seminars and workshops at which scientists, policy-makers and the public can debate crucial scientific matters.

Some hurdles must still be overcome, Zhu explains. Public participation in Chinese policy-making is a new concept to all involved. Government officials are not yet reconciled to having to justify political decisions to the populace, nor

are scientists used to explaining their research to a general audience. As a result, much of China's population is insufficiently informed about science and technology issues. "It's an important aspect of building a more democratic society," says Zhu. "It will come with time."

Yet people in China are eager to obtain more information and voice their views. In Shanghai, at one of several 'café scientifique' events organized by the British Council, stem-cell researcher Stephen Minger of King's College London and his Chinese colleagues had lively exchanges about stem-cell therapies with audiences of all ages and professions. Last month, the Beijing-based National Art Museum of China mounted an exhibition called *Synthetic Times*. Prominent installation artists from 29 countries explored issues such as identity, emotion, perception of reality, and the relationship between humans and technology in time and space.

"There is a lot of interest in science and technology from all sectors of the Chinese public," says Liu. "To channel that energy and curiosity properly is key to promoting the awareness of science and its social impact." ■

Jane Qiu is a science writer based in London and Beijing.
e-mail: jane@janeqiu.com

See Editorial, page 367.

A shared view of the heavens

A woodcut of Ferdinand Verbiest, the Kangxi Emperor's Flemish astronomer and mastermind of Beijing's Ancient Observatory, records a remarkable seventeenth-century cultural exchange. **Martin Kemp** explains.

Ferdinand Verbiest: Heaven on Earth
Museum of the History of Science
Oxford, UK
Until 7 September

CHINA

Not far from Beijing station, in a cityscape dominated by new buildings and multi-lane highways, stands a squat, ancient tower. On top sits the world's greatest historical ensemble of large-scale astronomical instruments. They were mainly designed and installed in 1673 by Ferdinand Verbiest, the Flemish Jesuit who was mathematician and astronomer to the Kangxi Emperor.

Verbiest makes a striking appearance in a coloured 1827 woodcut (pictured) by the Japanese artist, Utagawa Kuniyoshi. He stands in Chinese state robes, accompanied by smaller variants of his celestial globe and sextant while enumerating points on his fingers. The inscription on the print tells us that it portrays Chitasei Goyo, one of the 108 rogue heroes of the popular classical Chinese novel, *Water Margin*. How was the master of Chinese astronomy transformed into the clever strategist of a military gang?

The story begins with one of the most remarkable cultural exchanges of any era. It involves three successive Jesuit astronomers, sent as missionaries from Rome to China. Verbiest followed Matteo Ricci and Adam Schall von Bell to work as an astronomer at the emperor's court. The Jesuits were in fierce competition with traditional Chinese and Muslim astronomers for scientific and religious supremacy. At one stage, when the intellectual and political climate had moved against them, Schall and Verbiest were imprisoned under sentence of death by dismemberment, a gruesome fate they only narrowly avoided.

By 1699, Verbiest's star was in the ascendant with the emperor. He triumphed over his

BRITISH MUSEUM/BRIDGEMAN ART LIBRARY

Chinese rival in a contest to demonstrate the accuracy of his science, and reformed the Chinese calendar. A notable polymath and author, who designed cannon and steam-driven vehicles among other ingenious devices, Verbiest's most enduring achievement was the set of six new instruments for Beijing's observatory tower.

Taking inspiration from Tycho Brahe's ensemble of massive astronomy instruments on the Danish island of Hven, Verbiest spared no expense in establishing the world's definitive observatory. The great bronze celestial globe, for example, is almost 2 metres in diameter, and he boasted that it cost the massive sum of 50,000 taels, or silver pieces. Then, as now, astronomy was a costly science requiring big instruments.

To bring his achievements before the widest international audience, Verbiest published a set of 105 prints, mainly devoted to his observatory instruments and their manufacture, but also demonstrating Euclidian geometry, ballistics and various notable feats of engineering. The graphic technique of his illustrations exploits western-style draftsmanship for the instruments themselves, whereas the spaces within which they are located are drawn in the Chinese manner. Thus, the celestial globe is rendered in a convincingly plastic form, but the chequerboard tiling beneath it clearly does not observe the rules of linear perspective.

The intellectual traffic between China and Europe went both ways. The presence of the Jesuit scientists at the Chinese court led to a greater awareness in Europe of the richness of Chinese history, culture, science and technology. The thoughts of Confucius were made available to western philosophers when the first Latin edition of *Confucius, Philosopher of the Chinese* was published in Paris in 1686, prefaced by introductions to Chinese history, theology and the philosopher's own life.

The reach of Verbiest's fame, and of his splendid instruments, was considerable, as evinced by Kuniyoshi's spectacular print. And so to the question of why the artist cast a Japanese bandit hero in the guise of the famous astronomer. The answer probably lies in Goyo's Chinese name, Wu Yong, which means 'wise star'. Verbiest's instruments thus refer in a double sense to the hero's name and to his famed wisdom as a military strategist, which required expertise in maps, navigation and the various technologies over which Verbiest claimed mastery.

These interchanges between Jesuit and Chinese science and Japanese mythology remind us that global communication thrived long before our technological era. ■

Martin Kemp is research professor in the history of art at the University of Oxford, OX1 1PT, UK.

See Editorial, page 367.

are scientists used to explaining their research to a general audience. As a result, much of China's population is insufficiently informed about science and technology issues. "It's an important aspect of building a more democratic society," says Zhu. "It will come with time."

Yet people in China are eager to obtain more information and voice their views. In Shanghai, at one of several 'café scientifique' events organized by the British Council, stem-cell researcher Stephen Minger of King's College London and his Chinese colleagues had lively exchanges about stem-cell therapies with audiences of all ages and professions. Last month, the Beijing-based National Art Museum of China mounted an exhibition called *Synthetic Times*. Prominent installation artists from 29 countries explored issues such as identity, emotion, perception of reality, and the relationship between humans and technology in time and space.

"There is a lot of interest in science and technology from all sectors of the Chinese public," says Liu. "To channel that energy and curiosity properly is key to promoting the awareness of science and its social impact." ■

Jane Qiu is a science writer based in London and Beijing.
e-mail: jane@janeqiu.com

See Editorial, page 367.

A shared view of the heavens

A woodcut of Ferdinand Verbiest, the Kangxi Emperor's Flemish astronomer and mastermind of Beijing's Ancient Observatory, records a remarkable seventeenth-century cultural exchange. **Martin Kemp** explains.

Ferdinand Verbiest: Heaven on Earth
Museum of the History of Science
Oxford, UK
Until 7 September

CHINA

Not far from Beijing station, in a cityscape dominated by new buildings and multi-lane highways, stands a squat, ancient tower. On top sits the world's greatest historical ensemble of large-scale astronomical instruments. They were mainly designed and installed in 1673 by Ferdinand Verbiest, the Flemish Jesuit who was mathematician and astronomer to the Kangxi Emperor.

Verbiest makes a striking appearance in a coloured 1827 woodcut (pictured) by the Japanese artist, Utagawa Kuniyoshi. He stands in Chinese state robes, accompanied by smaller variants of his celestial globe and sextant while enumerating points on his fingers. The inscription on the print tells us that it portrays Chitasei Goyo, one of the 108 rogue heroes of the popular classical Chinese novel, *Water Margin*. How was the master of Chinese astronomy transformed into the clever strategist of a military gang?

The story begins with one of the most remarkable cultural exchanges of any era. It involves three successive Jesuit astronomers, sent as missionaries from Rome to China. Verbiest followed Matteo Ricci and Adam Schall von Bell to work as an astronomer at the emperor's court. The Jesuits were in fierce competition with traditional Chinese and Muslim astronomers for scientific and religious supremacy. At one stage, when the intellectual and political climate had moved against them, Schall and Verbiest were imprisoned under sentence of death by dismemberment, a gruesome fate they only narrowly avoided.

By 1699, Verbiest's star was in the ascendant with the emperor. He triumphed over his

BRITISH MUSEUM/BRIDGEMAN ART LIBRARY

Chinese rival in a contest to demonstrate the accuracy of his science, and reformed the Chinese calendar. A notable polymath and author, who designed cannon and steam-driven vehicles among other ingenious devices, Verbiest's most enduring achievement was the set of six new instruments for Beijing's observatory tower.

Taking inspiration from Tycho Brahe's ensemble of massive astronomy instruments on the Danish island of Hven, Verbiest spared no expense in establishing the world's definitive observatory. The great bronze celestial globe, for example, is almost 2 metres in diameter, and he boasted that it cost the massive sum of 50,000 taels, or silver pieces. Then, as now, astronomy was a costly science requiring big instruments.

To bring his achievements before the widest international audience, Verbiest published a set of 105 prints, mainly devoted to his observatory instruments and their manufacture, but also demonstrating Euclidian geometry, ballistics and various notable feats of engineering. The graphic technique of his illustrations exploits western-style draftsmanship for the instruments themselves, whereas the spaces within which they are located are drawn in the Chinese manner. Thus, the celestial globe is rendered in a convincingly plastic form, but the chequerboard tiling beneath it clearly does not observe the rules of linear perspective.

The intellectual traffic between China and Europe went both ways. The presence of the Jesuit scientists at the Chinese court led to a greater awareness in Europe of the richness of Chinese history, culture, science and technology. The thoughts of Confucius were made available to western philosophers when the first Latin edition of *Confucius, Philosopher of the Chinese* was published in Paris in 1686, prefaced by introductions to Chinese history, theology and the philosopher's own life.

The reach of Verbiest's fame, and of his splendid instruments, was considerable, as evinced by Kuniyoshi's spectacular print. And so to the question of why the artist cast a Japanese bandit hero in the guise of the famous astronomer. The answer probably lies in Goyo's Chinese name, Wu Yong, which means 'wise star'. Verbiest's instruments thus refer in a double sense to the hero's name and to his famed wisdom as a military strategist, which required expertise in maps, navigation and the various technologies over which Verbiest claimed mastery.

These interchanges between Jesuit and Chinese science and Japanese mythology remind us that global communication thrived long before our technological era. ■

Martin Kemp is research professor in the history of art at the University of Oxford, OX1 1PT, UK.

See Editorial, page 367.

Core caper

Journey to the Center of the Earth
Film directed by Eric Brevig
In UK and US cinemas now

S. RAYMOND/NEW LINE CINEMA

Jules Verne dismantled: *Journey to the Center of the Earth* is silly, but holds science in high regard.

When Jules Verne wrote *A Journey to the Centre of the Earth* in 1864, science was still coming to terms with the planet's extreme age, and Verne's story of a swiss-cheese globe containing vast seas and prehistoric creatures had a satisfying ring of plausibility. The novel's eccentric scientist, Otto Lidenbrock, invokes real-life researchers from Humphry Davy to Joseph Fourier, and the thrilling plot is regularly punctuated by scientific musings that were then cutting-edge.

The book may have inspired many to become geologists, but for recent generations of readers, the obvious impossibility of the subterranean voyage has detracted from its allure. It was even "too fantastic" for David Stevenson of the California Institute of Technology, Pasadena, who proposed an unmanned mission to probe Earth's core in this journal in 2003.

So this 2008 cinematic visit to Verne's strange subterranean world is more akin to fantasy than science fiction. *Journey to the Center of the Earth*, the new 3D film by special-effects guru Eric Brevig, is silly — in a good way. And within its imaginary world, the film holds science and fact in high regard.

The film is not an exact remake of the novel. Rather, it imagines a world where a few present-day maverick geologists called Vernites believe the novel to be fact not fiction. The action follows a geologist, played for broad comedy by Brendan Fraser, his sullen teenage nephew (Josh Hutcherson) and the Icelandic daughter of a missing Vernite (Anita Briem).

A sleight of hand with the science — a few "seismic readings" on a computer screen — gets the trio to the centre of Earth. But once underground, science saves the day as Fraser's character shows expert knowledge of mineral properties that rescues them from lava, dinosaurs and the like.

The film's tough scientist hero and its exciting caverns and formations might even have the effect on young audiences that the novel presumably had on previous proto-geologists. It vividly portrays the geological world of rocks and lava as diverse, dynamic and cool. It also pokes fun at the maverick scientist trope, with deadpan lines like "Although [he] was ridiculed by the scientific community, he was eventually found to be correct."

That said, the movie is pretty mindless. It has the standard comedic patter in the face of danger, with punchlines you can see coming all 6,400 kilometres from the centre of Earth. Mandatory shots take advantage of the 3D to make the audience jump. Happily, it doesn't take itself too seriously: "Eat your trilobite son; you've got to keep your strength up." And its new and improved 3D effects are a lot of fun to watch. ■

Emma Marris is a correspondent for *Nature*.

Geological history turned upside down

Worlds Before Adam: The Reconstruction of Geohistory in the Age of Reform
by Martin J. S. Rudwick
University of Chicago Press: 2008. 614 pp. $49.00

Geologists study Earth by applying principles borrowed from more fundamental sciences. Yet geology also has its own set of attitudes that have accrued during the discipline's long history. The nature of geological inquiry, involving a synthesis of historical and philosophical reasoning, lies at the heart of Martin Rudwick's fine new book.

Worlds Before Adam shows that the emergence of modern geology was comparable in its cultural impact with that of relativity or Darwinian evolution. Rudwick, an influential historian of Earth science, emphasizes geology's historical and causal approaches to understanding, complementing his magisterial book *Bursting the Limits of Time* (University of Chicago Press, 2005). This earlier work covered the period between 1787 and 1822, when French savant Georges Cuvier and his fellow continental geologists gave meaning to signs of the past, such as fossils and strata, in the same way as historians and archaeologists use monuments and archives to map human history. *Worlds Before Adam* looks at how the ideas generated by Cuvier and others came together with more theoretical concepts between 1820 and 1845.

Rudwick's books are myth-busters, of which writers of introductory geology texts and popularizations should take note. In both volumes he counters the Anglocentric view that James Hutton, William Smith and Charles Lyell were the founders of modern geology who shone their British intellectual light onto the darkness of continental musings. To a large degree, he argues, the reverse was the case.

Controversially, Rudwick challenges the view that geology's development is a story of secular progress. He shows that the founders of geology were almost all men of faith. Yet they often engaged in fierce debates with pseudo-scientists who ascribed absolute authority to readings of the Bible. Theologians have discredited such

Core caper

Journey to the Center of the Earth
Film directed by Eric Brevig
In UK and US cinemas now

S. RAYMOND/NEW LINE CINEMA

Jules Verne dismantled: *Journey to the Center of the Earth* is silly, but holds science in high regard.

When Jules Verne wrote *A Journey to the Centre of the Earth* in 1864, science was still coming to terms with the planet's extreme age, and Verne's story of a swiss-cheese globe containing vast seas and prehistoric creatures had a satisfying ring of plausibility. The novel's eccentric scientist, Otto Lidenbrock, invokes real-life researchers from Humphry Davy to Joseph Fourier, and the thrilling plot is regularly punctuated by scientific musings that were then cutting-edge.

The book may have inspired many to become geologists, but for recent generations of readers, the obvious impossibility of the subterranean voyage has detracted from its allure. It was even "too fantastic" for David Stevenson of the California Institute of Technology, Pasadena, who proposed an unmanned mission to probe Earth's core in this journal in 2003.

So this 2008 cinematic visit to Verne's strange subterranean world is more akin to fantasy than science fiction. *Journey to the Center of the Earth*, the new 3D film by special-effects guru Eric Brevig, is silly — in a good way. And within its imaginary world, the film holds science and fact in high regard.

The film is not an exact remake of the novel. Rather, it imagines a world where a few present-day maverick geologists called Vernites believe the novel to be fact not fiction. The action follows a geologist, played for broad comedy by Brendan Fraser, his sullen teenage nephew (Josh Hutcherson) and the Icelandic daughter of a missing Vernite (Anita Briem).

A sleight of hand with the science — a few "seismic readings" on a computer screen — gets the trio to the centre of Earth. But once underground, science saves the day as Fraser's character shows expert knowledge of mineral properties that rescues them from lava, dinosaurs and the like.

The film's tough scientist hero and its exciting caverns and formations might even have the effect on young audiences that the novel presumably had on previous proto-geologists. It vividly portrays the geological world of rocks and lava as diverse, dynamic and cool. It also pokes fun at the maverick scientist trope, with deadpan lines like "Although [he] was ridiculed by the scientific community, he was eventually found to be correct."

That said, the movie is pretty mindless. It has the standard comedic patter in the face of danger, with punchlines you can see coming all 6,400 kilometres from the centre of Earth. Mandatory shots take advantage of the 3D to make the audience jump. Happily, it doesn't take itself too seriously: "Eat your trilobite son; you've got to keep your strength up." And its new and improved 3D effects are a lot of fun to watch. ■

Emma Marris is a correspondent for *Nature*.

Geological history turned upside down

Worlds Before Adam: The Reconstruction of Geohistory in the Age of Reform
by Martin J. S. Rudwick
University of Chicago Press: 2008. 614 pp. $49.00

Geologists study Earth by applying principles borrowed from more fundamental sciences. Yet geology also has its own set of attitudes that have accrued during the discipline's long history. The nature of geological inquiry, involving a synthesis of historical and philosophical reasoning, lies at the heart of Martin Rudwick's fine new book.

Worlds Before Adam shows that the emergence of modern geology was comparable in its cultural impact with that of relativity or Darwinian evolution. Rudwick, an influential historian of Earth science, emphasizes geology's historical and causal approaches to understanding, complementing his magisterial book *Bursting the Limits of Time* (University of Chicago Press, 2005). This earlier work covered the period between 1787 and 1822, when French savant Georges Cuvier and his fellow continental geologists gave meaning to signs of the past, such as fossils and strata, in the same way as historians and archaeologists use monuments and archives to map human history. *Worlds Before Adam* looks at how the ideas generated by Cuvier and others came together with more theoretical concepts between 1820 and 1845.

Rudwick's books are myth-busters, of which writers of introductory geology texts and popularizations should take note. In both volumes he counters the Anglocentric view that James Hutton, William Smith and Charles Lyell were the founders of modern geology who shone their British intellectual light onto the darkness of continental musings. To a large degree, he argues, the reverse was the case.

Controversially, Rudwick challenges the view that geology's development is a story of secular progress. He shows that the founders of geology were almost all men of faith. Yet they often engaged in fierce debates with pseudo-scientists who ascribed absolute authority to readings of the Bible. Theologians have discredited such

views for centuries, but they still persist, with geologists continuing to refute them.

If contemporary lists of the greatest scientists feature a geologist at all, it is usually Lyell, a central figure in *Worlds Before Adam*. Lyell intended the title of his great multi-volume opus *Principles of Geology* (first published in 1830–1833) to recall Isaac Newton's *Principia*. He sought to recast geology on firm foundations, just as Newton had done for physics. Following his geologist contemporaries and predecessors, Lyell used the understanding of present-day causes to interpret the deep past — a principle termed actualism. Rudwick explains that Lyell's excellent descriptions of current geological processes, embellished with observations from his own geological excursions, derived from an original listing by the eighteenth-century German scholar Karl Ernst Adolf von Hoff. Lyell greatly extended the actualistic method by making pronouncements about how the complex geological processes of the past occurred through the progressive action of small-scale procedures that were still in operation, and by prescribing how geologists should reason about these past processes.

Rudwick shows that Lyell's ideas met with almost universal criticism. This was not caused by his advocacy of actualism, which was widely used, nor was any serious denunciation forthcoming from the biblical literalists, who were considered anti-scientific by Lyell and by his critics. Instead, the geological facts themselves seemed contrary to Lyell's vision of uniform action by small-scale processes operating over a long time. Examples include evidence for sudden mass extinctions from records in various 'bone caves', the existence of huge blocks sitting erratically out of geological place in the Alps and northern Europe, and deep U-shaped valleys containing streams too small to account for their excavation. Lyell's critics held that one should inquire into nature through evidence, rather than through privileged reasoning.

The great Cambridge polymath William Whewell named the two sides in the debate. Lyell's advocates he labelled 'uniformitarians'; their opponents he called 'catastrophists'. It is an irony of subsequent developments in geology, and a testimony to the success of Lyell's advocacy, that catastrophism came to be regarded as unconventional. This perverted Whewell's original intention, which was to show that the uniformitarians and Lyell were extreme in thinking that geologists should say in advance how nature works, through slow and uniform processes, before interpreting the evidence.

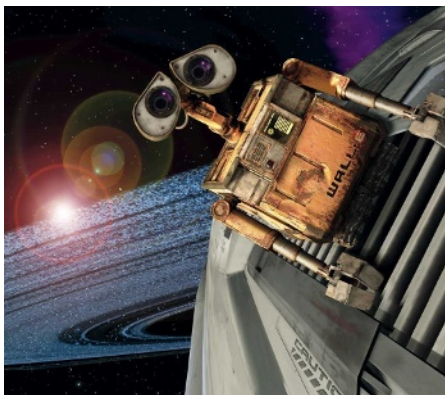
Worlds Before Adam concludes with the development of glacial theory, popularized in the nineteenth century by Cuvier's disciple Louis Agassiz, perhaps the greatest of the catastrophists. Agassiz's theory of the great spread of ice sheets during relatively recent geological time gained rapid acceptance among catastrophists because it accounted for many anomalous features originally ascribed to huge floods or tsunamis. However, Lyell resisted, remaining true to his epistemological project.

As we enter an era of global crises about water, energy and the environment, and as we seek to understand the development of our species among others in one corner of the Universe, geologists' perspectives give a means for both understanding and coping. In showing how these perspectives arose, Rudwick highlights an underappreciated, glorious advance in human thought, the documentation of which is a rather glorious achievement in itself. ■

Victor R. Baker is Regents' Professor of Hydrology and Water Resources, Planetary Sciences and Geosciences at the University of Arizona, Tucson, Arizona 85721, USA, and ex-president of the Geological Society of America.
e-mail: baker@hwr.arizona.edu

Romance among robots

WALL·E
Film directed by Andrew Stanton
In UK and US cinemas now



A few years ago, at the Massachusetts Institute of Technology's artificial intelligence lab, I met an android. Her conversation was perfunctory, mostly simple responses to my equally simple words, but her eyes, widening, narrowing and subtly changing angle, made a genuine emotional connection. That robot had me at "Hello". So it is with WALL·E (pictured), the eponymous hero of Disney–Pixar's new animated film. Part Mars rover, part Andy Hardy, WALL·E charms us every step of the way as he saves a planet while pursuing chaste robotic love.

The movie opens on a bleak future, reminiscent of films by director Ridley Scott at his dystopian best. Earth, abused and then abandoned by a population that never heeded Al Gore, lies grey and silent beneath the refuse of civilization. Punctuating the stillness is a small buzz of activity. WALL·E, a computerized rubbish compactor, perhaps descended from those robotic vacuum cleaners, dutifully pursues work he was programmed to do hundreds of years earlier, before humans gave up on the dream of refurbishing the planet. It's not a bad existence, but as he considers the oddments he has scavenged over time, particularly an old video tape of the film *Hello, Dolly!*, WALL·E recognizes that something is missing. That something soon materializes in the form of a robotic scout, sent to Earth to search for signs of photosynthesis. The scout is named EVE and … well, you can see where all this leads.

Movie buffs will enjoy *WALL·E*'s film references — from *2001: A Space Odyssey* (of course) to *Modern Times*. Science nerds will appreciate how both the story and the animation are informed by NASA and research into artificial intelligence. Pixar animators have mastered the literature on non-verbal communication; they have studied in detail the workings of robots from Mars rovers to assembly lines, and have internalized the stunning images from the Hubble and Spitzer space telescopes.

The animation in *WALL·E* is astonishing, but Pixar recognized long ago that technology alone does not fill cinemas. Stories do, and *WALL·E*'s creators are master storytellers. Sci-fi master Robert Heinlein maintained that there are only three plots in science fiction. All figure here: a sweet love story, the triumph of plucky stowaways over a power-hungry computer (remember HAL?), and a plea for planetary redemption. Moreover, the movie is funny. Eight-year-olds and octogenarians alike laughed throughout the screening, usually at the same time.

So, for animated sci-fi that honours both the science and the fiction, steal away to *WALL·E*. And, if you work at NASA's Jet Propulsion Laboratory in California, go and see it twice. When a future Mars rover angles its soulful head-lamps while asking for more funding, who at NASA will be able to refuse? ■

Andrew H. Knoll is Fisher Professor of Natural History at Harvard University, Cambridge, Massachusetts, and a member of NASA's Mars Exploration Rover science team.
e-mail: aknoll@fas.harvard.edu

views for centuries, but they still persist, with geologists continuing to refute them.

If contemporary lists of the greatest scientists feature a geologist at all, it is usually Lyell, a central figure in *Worlds Before Adam*. Lyell intended the title of his great multi-volume opus *Principles of Geology* (first published in 1830–1833) to recall Isaac Newton's *Principia*. He sought to recast geology on firm foundations, just as Newton had done for physics. Following his geologist contemporaries and predecessors, Lyell used the understanding of present-day causes to interpret the deep past — a principle termed actualism. Rudwick explains that Lyell's excellent descriptions of current geological processes, embellished with observations from his own geological excursions, derived from an original listing by the eighteenth-century German scholar Karl Ernst Adolf von Hoff. Lyell greatly extended the actualistic method by making pronouncements about how the complex geological processes of the past occurred through the progressive action of small-scale procedures that were still in operation, and by prescribing how geologists should reason about these past processes.

Rudwick shows that Lyell's ideas met with almost universal criticism. This was not caused by his advocacy of actualism, which was widely used, nor was any serious denunciation forthcoming from the biblical literalists, who were considered anti-scientific by Lyell and by his critics. Instead, the geological facts themselves seemed contrary to Lyell's vision of uniform action by small-scale processes operating over a long time. Examples include evidence for sudden mass extinctions from records in various 'bone caves', the existence of huge blocks sitting erratically out of geological place in the Alps and northern Europe, and deep U-shaped valleys containing streams too small to account for their excavation. Lyell's critics held that one should inquire into nature through evidence, rather than through privileged reasoning.

The great Cambridge polymath William Whewell named the two sides in the debate. Lyell's advocates he labelled 'uniformitarians'; their opponents he called 'catastrophists'. It is an irony of subsequent developments in geology, and a testimony to the success of Lyell's advocacy, that catastrophism came to be regarded as unconventional. This perverted Whewell's original intention, which was to show that the uniformitarians and Lyell were extreme in thinking that geologists should say in advance how nature works, through slow and uniform processes, before interpreting the evidence.

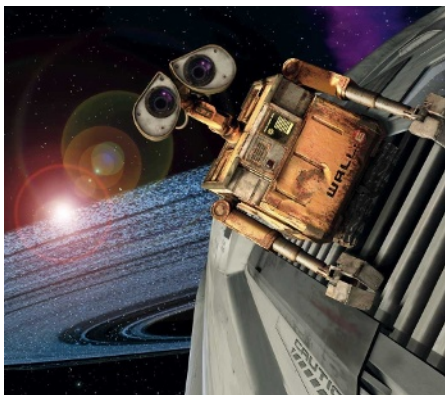
Worlds Before Adam concludes with the development of glacial theory, popularized in the nineteenth century by Cuvier's disciple Louis Agassiz, perhaps the greatest of the catastrophists. Agassiz's theory of the great spread of ice sheets during relatively recent geological time gained rapid acceptance among catastrophists because it accounted for many anomalous features originally ascribed to huge floods or tsunamis. However, Lyell resisted, remaining true to his epistemological project.

As we enter an era of global crises about water, energy and the environment, and as we seek to understand the development of our species among others in one corner of the Universe, geologists' perspectives give a means for both understanding and coping. In showing how these perspectives arose, Rudwick highlights an underappreciated, glorious advance in human thought, the documentation of which is a rather glorious achievement in itself. ■

Victor R. Baker is Regents' Professor of Hydrology and Water Resources, Planetary Sciences and Geosciences at the University of Arizona, Tucson, Arizona 85721, USA, and ex-president of the Geological Society of America.
e-mail: baker@hwr.arizona.edu

Romance among robots

WALL·E
Film directed by Andrew Stanton
In UK and US cinemas now



A few years ago, at the Massachusetts Institute of Technology's artificial intelligence lab, I met an android. Her conversation was perfunctory, mostly simple responses to my equally simple words, but her eyes, widening, narrowing and subtly changing angle, made a genuine emotional connection. That robot had me at "Hello". So it is with WALL·E (pictured), the eponymous hero of Disney–Pixar's new animated film. Part Mars rover, part Andy Hardy, WALL·E charms us every step of the way as he saves a planet while pursuing chaste robotic love.

The movie opens on a bleak future, reminiscent of films by director Ridley Scott at his dystopian best. Earth, abused and then abandoned by a population that never heeded Al Gore, lies grey and silent beneath the refuse of civilization. Punctuating the stillness is a small buzz of activity. WALL·E, a computerized rubbish compactor, perhaps descended from those robotic vacuum cleaners, dutifully pursues work he was programmed to do hundreds of years earlier, before humans gave up on the dream of refurbishing the planet. It's not a bad existence, but as he considers the oddments he has scavenged over time, particularly an old video tape of the film *Hello, Dolly!*, WALL·E recognizes that something is missing. That something soon materializes in the form of a robotic scout, sent to Earth to search for signs of photosynthesis. The scout is named EVE and … well, you can see where all this leads.

Movie buffs will enjoy *WALL·E*'s film references — from *2001: A Space Odyssey* (of course) to *Modern Times*. Science nerds will appreciate how both the story and the animation are informed by NASA and research into artificial intelligence. Pixar animators have mastered the literature on non-verbal communication; they have studied in detail the workings of robots from Mars rovers to assembly lines, and have internalized the stunning images from the Hubble and Spitzer space telescopes.

The animation in *WALL·E* is astonishing, but Pixar recognized long ago that technology alone does not fill cinemas. Stories do, and *WALL·E*'s creators are master storytellers. Sci-fi master Robert Heinlein maintained that there are only three plots in science fiction. All figure here: a sweet love story, the triumph of plucky stowaways over a power-hungry computer (remember HAL?), and a plea for planetary redemption. Moreover, the movie is funny. Eight-year-olds and octogenarians alike laughed throughout the screening, usually at the same time.

So, for animated sci-fi that honours both the science and the fiction, steal away to *WALL·E*. And, if you work at NASA's Jet Propulsion Laboratory in California, go and see it twice. When a future Mars rover angles its soulful head-lamps while asking for more funding, who at NASA will be able to refuse? ■

Andrew H. Knoll is Fisher Professor of Natural History at Harvard University, Cambridge, Massachusetts, and a member of NASA's Mars Exploration Rover science team.
e-mail: aknoll@fas.harvard.edu

Doctorate gets a lesson in management

Toward a Global PhD? Forces and Forms in Doctoral Education Worldwide
by Maresi Nerad and Mimi Heggelund
University of Washington Press: 2008.
320 pp. $30, £16.99

R. SPENCER/GETTY IMAGES

Doctoral studentships are more popular worldwide than ever despite their cost.

Higher education is increasingly market driven, but its most expensive product — the doctorate — is more popular than ever. Doctoral student numbers continue to rise as funding bodies support them with new initiatives.

Doctorates are expensive. For sponsors, spending around US$200,000 on a single individual can produce anxiety. For graduate students, the direct expenses are compounded by the cost of three or more years out of the labour market.

Given this investment, expectations of doctorates are increasing. This pressure is a thread through *Toward a Global PhD,* the product of a 2005 seminar convened in Seattle by the Center for Innovation and Research in Graduate Education at the University of Washington. The centre's director, Maresi Nerad, and Mimi Heggelund, its international coordinator, bring together 13 comprehensive reviews of doctoral education in 14 countries.

Nerad and Heggelund see doctoral education as the "primary source of research productivity and innovation in the global knowledge economy". They argue that it must both supply researchers and leaders in a wide range of occupations. Postgraduate training needs to instil skills alongside knowledge. European employers, for example, complain that doctoral graduates are too specialized and communicate poorly. To generate those who can "participate effectively in a corporate environment", collaboration must be promoted. International values must also be respected. Developing countries hope that the acquisition of such talents is not accompanied by a brain drain.

These varied needs must be addressed without compromising core skills or quality. A grounding in research methods remains key, along with knowledge of emerging areas such as ethics and intellectual property. International standards must be consistent and recognizable, despite vastly different resources. Candidates must also complete their studies and submit theses within a reasonable time.

Implementing these goals is difficult. The book's editors recognize that they only define parameters for future work, but make a noble attempt to draw conclusions by bringing together material that is not widely available. Each review is thorough and contains sufficient background on policy, and most analyse probable future developments. A key theme is the inequality of resources, numbers and research agendas. Other themes include the 'mode 2' team-based educational model developed by Michael Gibbons and others, brain drain, and the Bologna proposals for unifying doctoral education practices in the European Union.

The editors identify 15 future education trends, which reflect these and other themes including commoditization, the market economy, and the increasing use of English. Doctoral education, despite its growth, is now subject to more planning and direction than ever. The growth of formal evaluation mechanisms, supervision panels, explicitly stated admission standards and data collection are all examples. Yet such planning does not necessarily imply a narrowing focus. Doctorates are also becoming more international and more interdisciplinary, and some countries explicitly demand more diverse candidates.

Some chapters are too descriptive and view the problem disproportionately from the supply side, or are too conservative in identifying potential developments. In discussing globalization and brain drain, for example, the authors identify a trend for doctoral students to spend time out of their home country. More evidence of what works well would have been useful. Many sponsors and institutions have experimented with split-site doctorates, and it would be interesting to know how these vary in popularity, cost, completion rates and prestige.

Doctorates obtained by distance learning are also slowly increasing, yet this merits only one entry in the book's index. Further discussion of this would have been helpful, such as asking if distance learning provides a model for increasing supply while reducing unit cost, or if such opportunities improve equity by making high-quality supervision accessible to those unable to leave their jobs and countries. Other questions include whether distance learning helps to address brain drain issues, or if remote learners can form close bonds with supervisors.

The authors should also have scrutinized the economics of the PhD in more detail. They note that doctorates now seem to be aimed at solving specific challenges in developing industries rather than being pursued out of curiosity, for exploration or for love of a specific field. This has implications. Future education planners will need to become more familiar with their market — its needs, limitations and vulnerabilities. Experience suggests that, although demand for international education is increasing, the patterns of growth are uneven and complex. Given that 17 of the world's top universities are in the United States, this unevenness will be with us for some time to come.

In the best tradition of doctoral candidates, Maresi and Heggelund recognize the need for more research. Criticisms aside, *Toward a Global PhD* provides a useful framework for planners and providers alike. ■

John Kirkland is Deputy Secretary-General (Development) at the Association of Commonwealth Universities, 20–24 Tavistock Square, London WC1H 9HF, UK.
e-mail: j.kirkland@acu.ac.uk

ESSAY

The man who unveiled China

An English biochemist single-handedly changed the West's perception of China, revealing its past scientific glories and predicting more to come. **Simon Winchester** investigates the ongoing legacy of Joseph Needham.

CHINA

Seldom does the pork-and-rice reliability of a Chinese takeaway spring a revelation. One frigid December evening last year in Washington DC, as I was counting out money for a Shanghainese delivery man, I mentioned that I was writing a book about Joseph Needham, known in China as Li Yue-se. Hardly anyone in the United States knew of my chosen subject, so I was astonished when the delivery man gave a sudden and enthusiastic response. "How wonderful!" he replied. "Li Yue-se! The most famous Englishman ever to have lived in China! We are taught about him in school. He is loved in China, because he told us Chinese about ourselves. He helped us to feel proud about all we have done in the past."

The encounter underlined a sobering reality. Noel Joseph Terence Montgomery Needham (1900–1995), the sole architect and author of what is universally acknowledged to be the greatest and most authoritative of all books about China in the English language, is now far better known in the country about which he wrote than in his homeland where he wrote it.

This autumn marks 60 years since Needham began work on what would become his masterpiece: an enormous series of books entitled *Science and Civilisation in China*. Each volume was greeted with stunned admiration by critics and scholars around the world — the approbation growing as the number of volumes swelled. Together the books can now probably lay claim — although their author never did — to causing a profound mind-shift in the way the West came to regard the mysterious and long ill-regarded China.

NEEDHAM RES. INST.

Joseph Needham wrote nearly 3 million words on China before his death in 1995.

Fresh perspective

Needham was among the first in the Western world to realize the scope of China's historic achievements, and so to expect an equally glittering future for the nation. Now China is indeed rising to prominence again. And yet Needham, known and revered in China, is still very much a prophet without honour in the Western world. Except for within a scattering of academic centres, his name is little known or long forgotten.

Needham's interest in China came comparatively late. He read chemistry at the University of Cambridge, took a PhD in biochemistry, became a dedicated embryologist and married a woman in his department. He was a lifelong Marxist, a robust practising Christian, a nudist, a Morris dancer, an accordion player, a chain-smoker — and an ardent and unceasing womanizer. When a clever and pretty young Chinese scientist, Lu Gwei-djen, arrived in Cambridge from Nanjing in 1937 to study under Needham's wife Dorothy, he began a life-long affair. As he fell in love with Gwei-djen, he fell in love with her language and country.

When Needham began his flirtation, China was in a parlous state. This "booby nation", as the American poet and essayist Ralph Waldo Emerson had called it in 1824, was widely seen by Western mercantile classes as good for little more than the production of rhubarb, ceramics, silk and tea. As foreigners disdained it, so they also began to gnaw away at it. The British went to war over opium in 1839, seizing tracts of territory that included Hong Kong. The French followed suit in south China, the Germans in the east, the Russians and then the Japanese in the north. From 1937, Japan was engaged in a full-scale war.

As Needham was settling down to his calligraphy exercises in the studied calm of Cambridge, Tokyo's armies had occupied almost all of the eastern third of China. In one of the more

melancholy and little-known consequences of the invasion, many of the great eastern universities in cities such as Beijing, Shanghai, Tianjin and Nanjing were forced to flee westwards, to re-establish themselves as refugee colleges in safer mountain cities beyond the reach of the Japanese bombers. In 1941, these academics sent word to Britain that urgent help was needed to avoid the utter collapse of Chinese intellectual life.

The British government decided in 1942 to lend assistance. It chose Joseph Needham as a temporary diplomat to determine what was needed, arming him with a pistol in case of wartime emergencies. He arrived in Kunming in the spring of 1943, to begin three years of dauntingly difficult expeditions into the hitherto little-known heartlands of China.

Into the unknown

Lu Gwei-djen had advised her lover to keep an open mind while in China, and not to assume, as most Westerners did, that the country was an intellectual desert. Her advice struck home. On his very first day in Kunming, Needham records meeting an old gardener "in a little Mongol cap" top-grafting a plum tree. Needham's Mandarin proved good enough to communicate, and he soon learned that the man's grafting techniques were different from and far older than any similar techniques recorded in the West: the Chinese had, in effect, invented this aspect of plant husbandry.

A few days later, in a tailor's shop, Needham inquired into the antiquity of the Chinese abacus. He found it to be centuries older than Blaise Pascal's seventeenth-century calculating mechanism, the Arithmetique.

The findings set off a lifetime of inquiry in Needham's mind. He fast realized that China had created or developed scores of the ordinary, unsung underpinnings of human civilization. With painstaking precision, he began to list them.

Needham soon found Francis Bacon's long celebrated 'holy trinity' of gunpowder, printing and the compass to be Chinese. Also the stirrup, chains and chain drives, suspension bridges, blast furnaces, wheelbarrows, toilet paper, playing cards, inoculation, chess, the accurate establishment of π and much more. The Chinese, said Needham, had demonstrated "a promising start". He wrote: "The early Taoists (in China), not only curious about what they saw, but observing nature patiently and persistently, were [the world's] proto-scientists."

Although awed by all of this, Needham was also perplexed. "China was well endowed, and we can understand how it is that they obtained an early lead in technology," he wrote. "Yet it was Western Europe which discovered the method of developing the new natural science."

If the Chinese had invented so much, so early, why had their inventive energies dried up in the sixteenth century? Why was there never a Chinese Newton, a Kepler, Galileo or Einstein, when there had evidently in earlier times been a Chinese Euclid and Archimedes? He proposed what became known as the Needham question: why was almost all modern science the monopoly of the West?

The list of early Chinese achievements and this famous question became the twin underpinnings for a monumental book project that Needham began when he returned to Cambridge in 1948. Originally he did not think it would be such a huge task: his first letters to Cambridge University Press suggested he might compress his findings into a single volume. Maybe three, he wrote a few months later.

NEEDHAM RES. INST.

Lu Gwei-djen, Needham's inspiration and lover.

Then maybe seven, or eleven.

The first volume of *Science and Civilisation in China* was published in 1954, to general critical rapture. Volumes followed on a host of other topics. Whereas the first offered a general historical perspective, subsequent parts were more specific and included chemistry, nautics, astronomy, engineering, ceramics and botany. Most of them were enormous tomes, often split into several physical parts (none so large it could not be handily read in the bath, Needham insisted). At the time of Needham's death in 1995 he had completed 17 volumes, having written almost single-handedly more than three million words.

Readers sensed in the books something more than mere history. George Steiner, the polymathic critic, remarked in 1971 that like Proust, Needham had "made of remembrance both an act of moral justice and of high art".

The volumes, it was generally declared, had the power to change the world's mindset. Needham's books are littered with the famous Latinism *Ex oriente lux*, meaning light comes from the east. "It is time," he wrote, "that Christians realized that some of their highest values may be coming back to them from cultures and peoples far outside historical Christendom." Many agree that *Science and Civilisation in China* was instrumental in persuading the West's intellectual elite to shake themselves loose from their earlier preconceptions of moral, ethical and economic superiority — a legacy of paradigm-shifting that few other books can claim.

Mark Elvin, one of the greatest of contemporary scholars of Asian Pacific history, wrote in 1995 that after reading the books, "one's conception of the world has been transformed". A review in *Nature* of Needham's Volume 5, Part VII, which was largely devoted to the epic saga of the making of gunpowder, declared "no work of scholarship in the twentieth century has done as much to alter received ideas about the past" (W. H. McNeill *Nature* **326,** 751; 1987).

Few late-twentieth-century scholars of China who were made aware of the books — few scholars of any calling, in fact — long retained any dismissive notions of China. In time, and as international politics allowed for a greater travel to and correspondence with China, so these attitudes began to settle onto the wider world as a whole.

A difficult question

Needham deferred tackling his eponymous question until 1994, when he was more than 90 years old and working on what would be his final volume.

He first suggested that the lack of technological innovation was due in part to the lack of competition within the unified Imperial China under the Ming emperors. Europe, by contrast, was a mélange of competing and often warring states, their constant struggles for primacy leading to an endless tide of mercantile and military advances.

Needham also blamed what he called the 'bureaucratic feudalism' of late modern China. As late as the early twentieth century, it was the ultimate ambition of all clever young Chinese men to become not doctors or merchants or scientists but bureaucrats. The smartest men in the nation were bent on becoming officials who would run China just as it always had been run: unerringly true to the tenets of Confucius. Innovation, modernizing, reform — all these were inimical to the basic principles of those in charge.

The entry point into this bureaucratic system

for some 1,400 years was the legendarily rigorous Confucian examination system. The exam theoretically allowed for upward mobility throughout all classes, as success was down to merit rather than family or political connections. But it also promoted a fairly rigid type of learning: the exam required the faultless memorization of vast reams of classic Confucian texts. By the early 1900s, there was widespread antipathy against the archaic nature of the system. It was finally scrapped by the Qing imperial family in 1905, as part of a slew of reforms and modernizations they hoped would stave off their own downfall. They failed to save themselves. But in abolishing the examinations, and so freeing the brightest of the newly declared republic to strive for other goals, they triggered the birth of an entirely new Chinese attitude towards invention and innovation. Mass education with a Western-style curriculum was introduced and promoted.

Today, Chinese innovative energies have been fully unleashed once again. Universities are brimming with cash, brains and ambition (see page 382). There has been a noticeable increase in the number of patents applied for and papers published. The Chinese government recently passed a law stating, effectively, that it was acceptable for scientists to fail — an attempt to curb the traditional loss of face that long discouraged scientists whose experiments didn't work.

All the evidence suggests that a new golden age seems to be settling upon the country. True, for four recent centuries not a great deal of innovation went on. But this is a relatively brief period when ranged against the broad sweep of China's history. It seems that the original Needham question has been submerged by the rising tide of modern history.

BETTMANN/CORBIS

The Japanese invasion of China spurred universities there to call for help from abroad.

NEEDHAM RES. INST.

Needham rides into the remote areas near Dunhuang during his exploration of China's heartland.

None would today dare call China a 'booby nation'. The profound nature of the country's past achievement is now widely accepted, and seen as a natural precursor to the impressive future that China clearly has in store. Disdain has been replaced by awe: few doubt that China is on course to become a world force economically, intellectually and scientifically.

The work goes on

Needham's work continues. A research institute named after him in Cambridge has published eight volumes since his death in 1995, with three more in the works. The institute, which houses all of Needham's vast library, attracts scholars with as wide a variety of fascinations as Needham had himself: currently there are researchers investigating the role of millet in Neolithic China, sexuality and the rise of Chinese nationalism, Darwinism and modern China, early Chinese mathematics, the history of clocks, the history of Chinese hygiene, and much more. The presence of Needham is felt deeply here — an impressively large bust stands on a plinth beside the front door, and his ashes lie under a tree by the gravel pathway. The scholars are counselled to take note, as Needham would have wished, of the profound relevance of China's vast historical legacy, both to the country's condition today and to its future trajectory.

Back to the Chinese delivery man in Washington DC. By the most extraordinary coincidence, it turned out that I knew him. I realized that I had filmed him for a BBC documentary series in Shanghai in the 1980s, and had subsequently helped him get a visa and funds for a PhD in Philadelphia. He wanted to go and study in what he regarded as the world's leading nation: America.

He secured his degree, and by 1999 was living in Washington working on a secret communications protocol for the Department of Defense. After the terrorist attacks of 11 September 2001, he, along with all other non-citizens, was sacked. He was allowed to stay in the country as a consultant. Now, he told me that cold December night, his savings are all being put towards his new driving ambition: to go home.

His years in America had convinced him, unequivocally, of one overarching truth. America's time in the sun is rapidly coming to an end, and he should now return to the country poised to assume the world's leadership in its place: China. After telling me of this bold ambition, he added with pride and assurance that Li Yue-se — Joseph Needham — would have most readily and presciently agreed. ■

Simon Winchester is a writer living in Massachusetts. His book *The Man Who Loved China* was published in May.
e-mail: simonwinchester@mac.com

See Editorial, page 367.

ESSAY

The end of the science superpowers

Could the end of US world dominance over research mark the passing of national science giants, ask **J. Rogers Hollingsworth**, **Karl H. Müller** and **Ellen Jane Hollingsworth**.

From around 1735 until 1840 France led the world of science. This was the era of Antoine Lavoisier, Pierre-Simon Laplace and Claude Berthollet, with great advances in physics, mathematics, physiology and medicine. Centralization of the state and the education system in France, combined with a robust economy, made for a renowned science system. But ultimately, the centralized system led to rigidity and decline in the quality of science.

Next the nexus shifted to Germany, from the middle of the nineteenth century until the 1920s. This period saw the birth of a new type of research-oriented university, the creation of well-equipped laboratories, the emergence of numerous institutes, such as the Kaiser Wilhelm (later Max Planck) Institutes, and the growth of science-based industries such as dyes, pharmaceuticals and vaccines. In the first eleven years of the Nobel prizes, thirteen German scientists received awards in chemistry, medicine or physics — many more than any other country.

At the beginning of the twentieth century, the hub shifted to Britain. Over the next half century, scientific funding from government and industry rose, the university system was vigorous, and the country boasted numerous Nobel prizewinners: physicists Joseph John Thomson, the father and son team of William and Lawrence Bragg, Paul Dirac, James Chadwick and John Cockcroft; biologists Archibald Hill, Frederick Hopkins, Charles Sherrington, Edgar Adrian, Henry Dale and Howard Florey; and chemists Frederick Soddy and Alexander Todd. Then with the demise of the British Empire and the weakening of the British economy, this system of science declined too. The United States picked up the baton and holds it still.

The United States emerged from the Second World War as the world's economic superpower, facilitating the dominance of its system of science. Since then, American scientists have received more than half of the most prestigious awards in the sciences, such as Nobel, Lasker, Horwitz and Crafoord prizes. US researchers dominate scientific journals, accounting for more than 50% of the top 1% of cited papers and around 30% of all published papers. The United States also attracts talented young scientists for advanced training, echoing the migration of thousands of Americans to German universities during the second half of the nineteenth century and the later flow to Britain of scientists from across the British Empire.

Yet history suggests that the United States has no cause for complacency. Patterns in the rise and fall of former leading scientific nations imply that, unless serious steps are taken, the United States could look back on the early twenty-first century as the peak of its scientific dominance. Each former giant of science emerged when the society's economy became extraordinarily robust by world standards. As the French, German and British economies declined relative to the world's most dynamic centres of fiscal growth, so did their science systems. The independence and flexibility that once characterized their research systems diminished markedly. Each former scientific power, especially during the initial stages of decline, had the illusion that its system was performing better than it was, overestimating its strength and underestimating innovation elsewhere. The elite could not imagine that the centre would shift.

Meanwhile, fundamental changes over the past few decades in economics, funding, communication, organizational structure, and specialization could mean that the United States is not simply poised to cede its scientific throne to a national successor such as China. Rather, the end of America's era as the scientific hegemon could also be the end of the era of scientific hegemons.

> "The decline of the US economy is facilitating the strengthening of science elsewhere."

State of the union

Since 1945, the number of scientific papers and journals in highly industrialized societies — particularly the United States — has risen almost exponentially, while the proportion of the workforce in research and development and the percentage of gross national product devoted to it have grown more modestly. Yet the rate at which truly creative work emerges has remained relatively constant. In terms of the scale of research efforts to make major scientific breakthroughs, there are diminishing returns.

Americans have led the way in the emergence of 'big science', with, for example, the Manhattan Project, the Jet Propulsion Lab and the Lawrence Livermore, Argonne and Brookhaven national laboratories. Indeed, in all fields there has been a shift to collective research. One of the virtues of large-scale science is the ability to organize sizeable groups with different skills, ideas and resources. Teams produce many more papers than individuals do, leading to the boom in science publishing. In recent decades, the number of authors per paper has more than doubled. Moreover, team-authored papers are 6.3 times more likely to receive at least 1,000 citations (S. Wuchty, B. F. Jones and B. Uzzi *Science* **316,** 1036–1039; 2007).

In some fields, this transformation towards big science has built in irreversible constraints. During the past half century, universities, research institutes and pharmaceutical companies have swelled in number. Many universities have become increasingly bureaucratic and fragmented, with huge departments constructed like silos. As a result, many scientists have considerable difficulty in communicating across fields. The number of scientists, postdocs, research assistants, technicians and secretaries has mushroomed. To manage large scientific organizations, multiple levels of management have developed, with leaders of subgroups, chairs of departments, associate deans, deans of colleges, provosts for academic affairs, chancellors and vice-presidents for research, for business affairs and for legal affairs.

In some respects, the research segments of many US universities have become like holding companies. As long as researchers can bring in large research grants and pay substantial institutional overhead costs, universities are happy to have the income. Granting agencies and universities, realizing that this kind of structure has become dysfunctional, have made serious efforts to reduce the number of managerial levels and to develop matrix-type teams to minimize organizational rigidities. However, organizational inertia hampers these efforts.

With the ballooning of publications, universities, funding agencies and reviewers have less time to evaluate scientific papers carefully, and rely more and more on quantitative measures based on citation statistics. Scientists are increasingly assessed by the number of papers


D. MACKIE

they have authored. At the same time, the increasing commercialization of science has tended to emphasize short-term scientific horizons. All these factors threaten the future quality of American science.

Altering the dynamics

If funding agencies and leaders in the scientific community emphasize commercialization of science in large-scale research environments, the US system risks losing its flexibility and its capacity to make major fundamental discoveries as bases for new applications some 40 or 50 years hence. Often, knowledge for major discoveries is created during a largely unanticipated and unplanned stage of research and produces various unintended consequences.

Excellence in science requires nimble, autonomous organizations — qualities more likely to be found in small research settings. Dozens of scientists who made significant advances did so in organizations with fewer than 50 full-time researchers. In the recent past, some of the most creative small centres were the Rockefeller University in New York, the Salk Institute in San Diego, California, the Basel Institute for Immunology in Switzerland, the Laboratory of Molecular Biology in Cambridge, UK, and various Max Planck Institutes in Germany. In the past decade Nobel prizes have been awarded to scientists for work done in relatively small settings: Günter Blobel (physiology or medicine), Ahmed Zewail (chemistry), Paul Greengard (physiology or medicine), Andrew Fire (physiology or medicine), Roderick MacKinnon (chemistry) and Gerhard Ertl (chemistry).

America's science system could enhance its performance by creating several dozen small research organizations in interdisciplinary domains or in emerging fields, modelled along the lines of the organizations mentioned above. In recent years, there have been several such efforts — the Howard Hughes Medical Institute's Janelia Farm in Chevy Chase, Maryland, the Santa Fe Institute in New Mexico, the Institute Para Limes in Warnsveld, the Netherlands, and the new Institute for Quantum Optics, Quantum Nanophysics and Quantum Information in Vienna.

Last of the giants?

The decline of the US economy relative to those of the rest of the world is facilitating the strengthening of science elsewhere. An evolving multi-polar world economy is leading to multiple centres of science — the United States, the European Union, Japan, China, Russia and possibly India. The increasing wealth of several of these societies is enabling them to lure back many younger scientists trained abroad in the world's leading institutions.

A remarkable aspect of this change has been the rapidity with which China has emerged as an important science power. For example, China was fourteenth in the world in production of science and engineering papers in 1995; by 2005, as the Chinese economy boomed, it was fifth in the production of papers, according to Thomson Reuters ISI. By 2007 it was second. Between 1985 and 2005, the number of natural sciences and engineering doctoral degrees awarded in China increased sevenfold, so that by 2005 China was third in the world. Moreover, in recent years more and more senior expatriates have been returning to China.

The mobility of researchers and their funds across continents is rising rapidly: Europeans are moving in larger numbers to Asia and vice versa. Leading journals with articles from more countries play a major part in the governance of scientific practices and the coordination of scientists across the globe. Overall, there is more uniformity in methodology, training and publication, with open-access publishing and the Internet contributing to a globalized science system.

All in all, it seems unlikely that we will witness another unrivalled scientific behemoth in the mould of France, Germany, Britain and the United States. ■

J. Rogers Hollingsworth is professor of history at the University of Wisconsin (Madison), 455 North Park Street, Madison, Wisconsin 53706, USA.
e-mail: hollingsjr@aol.com
Karl H. Müller is director of the Vienna Institute for Social Science Documentation and Methodology (WISDOM), Maria Theresienstrasse 9/5, A-109 Vienna, Austria.
Ellen Jane Hollingsworth is senior scientist at the University of Wisconsin (Madison), 455 North Park Street, Madison, Wisconsin 53706, USA.

See Editorial, page 367, and News Feature, page 382. See also http://tinyurl.com/64r88t for further reading.

NEWS & VIEWS

I. BERRY/MAGNUM PHOTOS

Arsenic alert. The red paint on this well-head in Jhikargachha, Bangladesh, is a warning that the water is impure. But there is often no choice of water source.

ENVIRONMENTAL SCIENCE

Poisoned waters traced to source

Charles F. Harvey

South Asia's well-water is widely polluted with arsenic, but no one has located the source. A study on the Mekong River finds that contamination begins in pond sediments, and is spread by groundwater flow to wells.

Millions of people living in the Ganges delta of Bangladesh and West Bengal drink groundwater contaminated by arsenic. Many more ingest arsenic-contaminated water drawn from wells along the Mekong and Red rivers of Cambodia and Vietnam, and probably also along the Irrawaddy River in Myanmar. Arsenic dissolves in groundwater from naturally deposited sediments, but its localization within sedimentary aquifers is puzzling — wells from which severely contaminated water is drawn can be found a mere ten metres away from safe wells. On page 505 of this issue, Polizzotto *et al.*[1] describe the first study that traces the contamination from its source. The authors show that, at a site near the Mekong River, arsenic-laden groundwater originates from ponds. This water flows horizontally through the aquifer, beneath groundwater that percolates through soils, to contaminate downstream wells.

To understand how contaminated water may be traced back to its source, one must consider the physics of solute transport by groundwater flowing through aquifers. Groundwater flow-paths through typical undeveloped aquifers are arranged in a largely horizontal pattern (Fig. 1, overleaf), in which successively deeper layers originate from sources farther away[2]. Little mixing occurs across these flow-paths[3], so solute concentrations within an aquifer are also layered, with each deeper layer representing the biogeochemical outcome of water inputs from more distant sources.

By applying an understanding of these physical processes, Polizzotto *et al.*[1] arrived at two crucial insights. First, they realized that if they selected a location where groundwater flow is primarily perpendicular to the Mekong River, a single transect of wells of different depths aligned with that flow would sample water from the various layers of the subsurface system, fully mapping the local three-dimensional system. Second, they recognized that the chemical composition of groundwater at different depths reflects the different origins of the water — thus avoiding the common mistake of assuming that the composition is purely determined by a gradual process of change as solutes react with aquifer sediments over time. Although it is true that deeper groundwater is usually older, the water found at shallow depths does not flow straight down to deeper levels. By analysing the arsenic content of samples from different depths and locations at their site, Polizzotto *et al.* conclude that the local arsenic contamination originates from nearby pond sediments. So how can this source be explained?

The arsenic originally comes from eroded Himalayan sediments that have washed down into low-lying regions. It is widely believed that this arsenic dissolves and enters the groundwater under anaerobic conditions. It is therefore unsurprising to find that highly contaminated groundwater originates from pond sediments: the steady settling and decomposition of organic material at the bottom of tropical ponds takes up all the oxygen that diffuses, or is carried by downward flow, into the sediment.

Water passing through pond sediments could

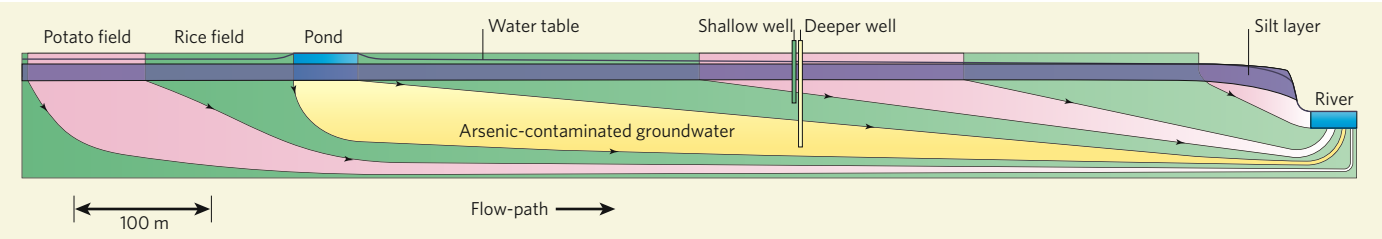


Figure 1 | Arsenic contamination in south Asian aquifers. Polizzotto *et al.*[1] report that arsenic contamination of groundwater at a site on the Mekong delta can be traced to ponds that supply the aquifer with water. The diagram shows groundwater flow through the cross-section of a system similar to that studied by the authors. Groundwater from different sources (potato fields, rice fields and a pond) flows in layers towards the river. A shallow well 'downstream' from the pond accesses water that started out from a rice field. A deeper well draws contaminated water that originated from the pond.

also contain organic carbon that, on decomposition, might help liberate arsenic from deeper sediments, adding to the contamination. But any organic carbon that is already contained in deeper aquifer sediments probably contributes less to biogeochemical processes because it is not replenished, and what remains is typically of low reactivity. Polizzotto and colleagues used carbon dating to show that the inorganic carbon dissolved in contaminated water at their site is young, as would be expected if it comes from pond sediments.

The authors' results raise the question of whether similar processes are responsible for the arsenic contamination observed in other aquifers throughout south Asia. If so, wells could be placed so as to avoid drawing groundwater that originates from ponds or similar bodies. Answering this question requires fieldwork at other sites, but several observations support the idea that the proposed mechanism[1] for arsenic contamination occurs elsewhere in south Asia. For example, natural and man-made ponds are ubiquitous in the region. Because the pond sediments are always saturated with water, they are more likely to be anoxic than soils — which, even under rice cultivation, are exposed to air several times a year.

Another clue is that arsenic concentrations in groundwater often increase with depth. This might also be indicative of surface-water inputs. Wells are never installed in ponds, so well-water drawn from the top of the aquifer will have passed only through the surrounding soil, rather than through anoxic sediments that would cause arsenic contamination. Deeper wells access groundwater that might have originated from ponds or similar bodies, and would therefore have been contaminated by arsenic.

The challenge in the Ganges delta, where the largest population of people is exposed to arsenic contamination, is that groundwater flow is exceedingly complex, depending on both geographical and temporal factors. Humans complicate the picture further — irrigation pumping drives groundwater in three-dimensional patterns that shift with the seasonal monsoon cycle. Furthermore, the widespread excavation of ponds and the explosive growth of irrigation pumping are changing subsurface solute concentrations over the course of decades[4].

In the United States and Europe, groundwater contamination sites are often first studied by characterizing groundwater flow. But in south Asia, research has focused on the biogeochemistry of arsenic transfer from aquifer substrate materials. This difference is understandable, because the source of dissolved arsenic in south Asia is naturally occurring sediments, rather than contaminant spills as in the United States and Europe. But Polizzotto and colleagues' work[1] clearly demonstrates that groundwater flow can control the localization of arsenic that originates from sediments. Understanding groundwater movement at more complex sites than the Mekong will require in-depth characterization of physical hydrogeology, such as that routinely used for small contamination sites in the United States and Europe. Although costly, this would surely be a small price to pay for the benefit of securing safe, clean drinking water for millions. ■

Charles F. Harvey is in the Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA.
e-mail: charvey@mit.edu

1. Polizzotto, M. L., Kocar, B. D., Benner, S. G., Sampson, M. & Fendorf, S. *Nature* **454,** 505–508 (2008).
2. Freeze, R. A. & Cherry, J. A. *Groundwater* (Prentice Hall, 1979).
3. Gelhar, L. W. *Stochastic Subsurface Hydrology* (Prentice Hall, 1993).
4. Harvey, C. F. *et al. Chem. Geol.* **228,** 112–136 (2006).

PHYSIOLOGY

Myoglobin's new clothes

Andrew Cossins and Michael Berenbrink

Nitric oxide generated from the nitrite ion limits the tissue damage caused by restricted blood flow. Gene knockout experiments in mice now reveal that myoglobin is the mediator of this effect.

All students of biology encounter the richly pigmented protein myoglobin early in their education, where it provides the first and most famous example of a revealed protein structure combined with a straightforward physiological role. Its restricted distribution to endurance muscle and heart cells throughout the vertebrates, and its notable expression in diving mammals, are a reflection of its widely accepted function in cellular oxygen transport and oxygen buffering. That function is to support high levels of aerobic muscular activity, and to deal out stored oxygen during extended periods of hypoxic — low oxygen — physiological conditions.

Given that all mammals have large quantities of myoglobin, experimental deletion of the genes concerned should have severe effects. Imagine, then, the surprise when the first myoglobin-knockout mouse[1] seemed to be perfectly normal, with no obvious ill effects during exercise and hypoxia. Unease about the accepted dogma has grown into debate about alternative functions for myoglobin[2–5], and has culminated in work by Hendgen-Cotta *et al.*, just published in *Proceedings of the National Academy of Sciences*[6]. Their paper shows that myoglobin knockout in mice has consequences that are both surprising and medically important.

This story has its origins in the apparently ordinary inorganic anion, nitrite (NO_2^-). Sodium nitrite is an ancient means of curing meat, but it is also used to maintain the desirable cherry-red colour of meat while it is on supermarket shelves. It has been used medicinally in high concentrations as a vaso- and bronchodilator (that is, for dilating blood vessels and airways), and as a treatment for cyanide poisoning, and the organic derivative amyl nitrite is used recreationally as a psychoactive drug. Environmental nitrite has been identified as an aquatic pollutant[7], and nitrite derived from contaminated drinking water is the cause of blue-baby syndrome[8]. This is an ailment in which human infants suffer from nitrite-induced formation of methaemoglobin,

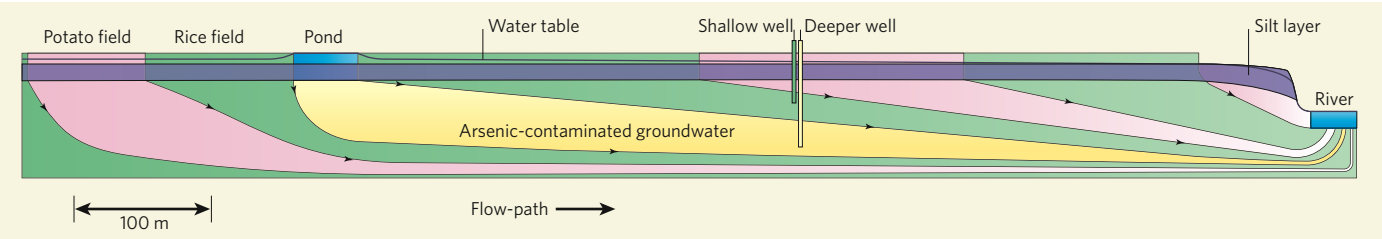


Figure 1 | Arsenic contamination in south Asian aquifers. Polizzotto *et al.*[1] report that arsenic contamination of groundwater at a site on the Mekong delta can be traced to ponds that supply the aquifer with water. The diagram shows groundwater flow through the cross-section of a system similar to that studied by the authors. Groundwater from different sources (potato fields, rice fields and a pond) flows in layers towards the river. A shallow well 'downstream' from the pond accesses water that started out from a rice field. A deeper well draws contaminated water that originated from the pond.

also contain organic carbon that, on decomposition, might help liberate arsenic from deeper sediments, adding to the contamination. But any organic carbon that is already contained in deeper aquifer sediments probably contributes less to biogeochemical processes because it is not replenished, and what remains is typically of low reactivity. Polizzotto and colleagues used carbon dating to show that the inorganic carbon dissolved in contaminated water at their site is young, as would be expected if it comes from pond sediments.

The authors' results raise the question of whether similar processes are responsible for the arsenic contamination observed in other aquifers throughout south Asia. If so, wells could be placed so as to avoid drawing groundwater that originates from ponds or similar bodies. Answering this question requires fieldwork at other sites, but several observations support the idea that the proposed mechanism[1] for arsenic contamination occurs elsewhere in south Asia. For example, natural and man-made ponds are ubiquitous in the region. Because the pond sediments are always saturated with water, they are more likely to be anoxic than soils — which, even under rice cultivation, are exposed to air several times a year.

Another clue is that arsenic concentrations in groundwater often increase with depth. This might also be indicative of surface-water inputs. Wells are never installed in ponds, so well-water drawn from the top of the aquifer will have passed only through the surrounding soil, rather than through anoxic sediments that would cause arsenic contamination. Deeper wells access groundwater that might have originated from ponds or similar bodies, and would therefore have been contaminated by arsenic.

The challenge in the Ganges delta, where the largest population of people is exposed to arsenic contamination, is that groundwater flow is exceedingly complex, depending on both geographical and temporal factors. Humans complicate the picture further — irrigation pumping drives groundwater in three-dimensional patterns that shift with the seasonal monsoon cycle. Furthermore, the widespread excavation of ponds and the explosive growth of irrigation pumping are changing subsurface solute concentrations over the course of decades[4].

In the United States and Europe, groundwater contamination sites are often first studied by characterizing groundwater flow. But in south Asia, research has focused on the biogeochemistry of arsenic transfer from aquifer substrate materials. This difference is understandable, because the source of dissolved arsenic in south Asia is naturally occurring sediments, rather than contaminant spills as in the United States and Europe. But Polizzotto and colleagues' work[1] clearly demonstrates that groundwater flow can control the localization of arsenic that originates from sediments. Understanding groundwater movement at more complex sites than the Mekong will require in-depth characterization of physical hydrogeology, such as that routinely used for small contamination sites in the United States and Europe. Although costly, this would surely be a small price to pay for the benefit of securing safe, clean drinking water for millions. ■

Charles F. Harvey is in the Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA.
e-mail: charvey@mit.edu

1. Polizzotto, M. L., Kocar, B. D., Benner, S. G., Sampson, M. & Fendorf, S. *Nature* **454,** 505–508 (2008).
2. Freeze, R. A. & Cherry, J. A. *Groundwater* (Prentice Hall, 1979).
3. Gelhar, L. W. *Stochastic Subsurface Hydrology* (Prentice Hall, 1993).
4. Harvey, C. F. *et al. Chem. Geol.* **228,** 112–136 (2006).

PHYSIOLOGY

Myoglobin's new clothes

Andrew Cossins and Michael Berenbrink

Nitric oxide generated from the nitrite ion limits the tissue damage caused by restricted blood flow. Gene knockout experiments in mice now reveal that myoglobin is the mediator of this effect.

All students of biology encounter the richly pigmented protein myoglobin early in their education, where it provides the first and most famous example of a revealed protein structure combined with a straightforward physiological role. Its restricted distribution to endurance muscle and heart cells throughout the vertebrates, and its notable expression in diving mammals, are a reflection of its widely accepted function in cellular oxygen transport and oxygen buffering. That function is to support high levels of aerobic muscular activity, and to deal out stored oxygen during extended periods of hypoxic — low oxygen — physiological conditions.

Given that all mammals have large quantities of myoglobin, experimental deletion of the genes concerned should have severe effects. Imagine, then, the surprise when the first myoglobin-knockout mouse[1] seemed to be perfectly normal, with no obvious ill effects during exercise and hypoxia. Unease about the accepted dogma has grown into debate about alternative functions for myoglobin[2–5], and has culminated in work by Hendgen-Cotta *et al.*, just published in *Proceedings of the National Academy of Sciences*[6]. Their paper shows that myoglobin knockout in mice has consequences that are both surprising and medically important.

This story has its origins in the apparently ordinary inorganic anion, nitrite (NO_2^-). Sodium nitrite is an ancient means of curing meat, but it is also used to maintain the desirable cherry-red colour of meat while it is on supermarket shelves. It has been used medicinally in high concentrations as a vaso- and bronchodilator (that is, for dilating blood vessels and airways), and as a treatment for cyanide poisoning, and the organic derivative amyl nitrite is used recreationally as a psychoactive drug. Environmental nitrite has been identified as an aquatic pollutant[7], and nitrite derived from contaminated drinking water is the cause of blue-baby syndrome[8]. This is an ailment in which human infants suffer from nitrite-induced formation of methaemoglobin,

an inactive form of haemoglobin, the other famed member of the globin family.

Surprisingly low concentrations of nitrite, near the physiological level, change blood-flow properties and may modulate blood pressure[9]. Low levels of nitrite also act to suppress the activity of mitochondria, the organelles that are the main source of energy generation in a cell. In so doing, nitrite reduces the generation of mitochondrion-derived damaging reactive oxygen species (ROS)[2] in tissues suffering from ischaemia and from reperfusion damage. These conditions, respectively, are an interruption of blood flow and oxygen delivery to tissues caused by blocked blood vessels, and damage caused by a surge in the production of ROS.

Understanding how nitrite exerts its beneficial effects is of obvious physiological and medical significance[10]. Those effects are mainly linked to its reductive conversion to nitric oxide (NO), which relaxes the smooth-muscle lining of arterioles, the intermediate blood vessels between arteries and capillaries, and so dilates them[11]. But nitric oxide also acts on heart muscle, where it decreases muscle contractility and heart-beat rate by binding to mitochondrial proteins and suppressing the electron-transport chain that is essential in energy production[12]. This reduces the demand for oxygen and the ROS-induced damage during times of local oxygen deficiency. Nitric oxide synthases are generally regarded as the source of arteriolar nitric oxide by metabolizing the amino acid L-arginine. But there is increasing evidence that tissue nitric oxide can also be generated from nitrite through a nitrite reductase activity mediated by proteins such as xanthine oxidoreductase, deoxyhaemoglobin and deoxymyoglobin, or by a non-enzymatic reaction under acidic conditions[13].

The hypothesis that haemoglobin and myoglobin can mediate production of nitric oxide has been particularly controversial, because these molecules had been regarded as dioxygenases whose principal function is to scavenge surplus nitric oxide[4]. Hendgen-Cotta *et al.*[6], however, now clarify matters (Fig. 1). With myoglobin-knockout mice, they provide evidence that nitrite is the source of nitric oxide that protects heart cells following an episode of ischaemia or reperfusion, and — crucially — that deoxymyoglobin acts as the necessary nitrite reductase.

Thus, mice without myoglobin cannot generate nitrite-dependent nitric oxide, or nitrosylate heart muscle, or generate the cyclic GMP that is the essential signalling agent in producing vasodilation. Furthermore, compared with normal mice, the hearts of the knockout mice do not recover from experimentally imposed ischaemia; and these mice show no evidence of nitrite-induced reduction in the damage to heart tissue caused by blood-vessel blockage. On this evidence, then, myoglobin is both a reductase and a dioxygenase — it respectively produces and scavenges nitric oxide in deoxygenated and oxygenated conditions.

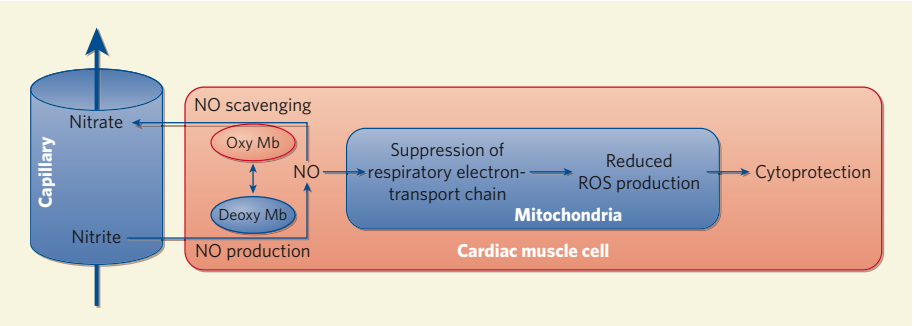


Figure 1 | Myoglobin as a reductase and an oxygenase. Myoglobin (Mb) respectively produces and scavenges nitric oxide (NO) in deoxygenated and oxygenated conditions. In its action as a reductase[6], deoxygenated myoglobin generates NO from circulating nitrite. This process is activated in cardiac muscle cells under hypoxic stress, where it suppresses the production of damaging reactive oxygen species (ROS) in mitochondria, so protecting the muscle cells from damage. Excess NO is reconverted to nitrate by oxymyoglobin acting as a dioxygenase.

The haem-containing globin superfamily includes two of the most-studied proteins in history, so it is astonishing to find that one of them has hitherto unidentified functions. But we can be confident that there will be more surprises, for at least two reasons. Myoglobin has been identified in tissues such as brain and liver[14], suggesting that its versatility in function applies to more than just muscle. And no role has yet been established for some other members of the superfamily, such as cytoglobin and neuroglobin[15]. ■

Andrew Cossins and Michael Berenbrink are in the School of Biological Sciences, University of Liverpool, Liverpool L69 7ZB, UK.
e-mail: cossins@liverpool.ac.uk

1. Garry, D. J. *et al. Nature* **395,** 905–908 (1998).
2. Shiva, S. *et al. Circ. Res.* **100,** 654–661 (2007).
3. Rayner, B. S., Wu, B.-J., Raftery, M., Stocker, R. & Witting, P. K. *J. Biol. Chem.* **280,** 9985–9993 (2005).
4. Brunori, M. *Trends Biochem. Sci.* **26,** 209–210 (2001).
5. Flögel, U. *et al. Proc. Natl Acad. Sci. USA* **98,** 735–740 (2001).
6. Hendgen-Cotta, U. B. *et al. Proc. Natl Acad. Sci. USA* doi:10.1073/pnas.0801336105 (2008).
7. Jensen, F. B. *Comp. Biochem. Physiol. Mol. Integr. Physiol.* **135,** 9–24 (2003).
8. Powlson, D. S. *et al. J. Environ. Qual.* **37,** 291–295 (2008).
9. Cosby, K. *et al. Nature Med.* **9,** 1498–1505 (2003).
10. Kumar, D. *et al. Proc. Natl Acad. Sci. USA* **105,** 7540–7545 (2008).
11. Furchgott, R. F. *Angew. Chem. Int. Edn* **38,** 1870–1880 (1999).
12. Rassaf, T. *et al. Circ. Res.* **100,** 1749–1754 (2007).
13. Lundberg, J. O., Weitzberg, E. & Gladwin, M. T. *Nature Rev. Drug Discov.* **7,** 156–167 (2008).
14. Fraser, J. *et al. Proc. Natl Acad. Sci. USA* **103,** 2977–2981 (2006).
15. Hankeln, T. *et al. J. Inorg. Biochem.* **99,** 110–119 (2005).

MOLECULAR COMPUTING

A layer of logic

A. Prasanna de Silva

Silicon chips have thousands of electronic logic gates etched on them. But there are other ways to decorate monolithic surfaces with logic gates, as a system using metal complexes secured to glass slides shows.

Logic lies at the heart of modern computers[1], and the components that carry out its operations are logic gates. On the basis of a number of digital inputs (the presence or absence of some condition conventionally represented as 1's or 0's), a logic gate produces a single digital output. For example, a simple AND gate produces a 1 output when both of its inputs are 1, and a 0 output if either or both inputs are 0. In electronic computers, logic gates are sculpted on the surface of silicon wafers. Their inputs and outputs are electrical voltages, but these are not the only systems that can form logic circuits. Writing in *Angewandte Chemie*, Gupta and van der Boom[2] describe a series of redox-active molecules anchored on glass surfaces that function as chemical logic gates.

The practice of using molecules as logic gates is about 15 years old[3], and could be said to be entering a turbulent adolescence. Although electronic logic circuits have been highly successful, molecular logic gates have some potential advantages — not least that they can operate in much smaller spaces. Molecular logic gates use concentrations of specific chemicals as their inputs, and the production of a particular chemical species, often accompanied by a colour change or similarly easy-to-identify property, as their outputs. So far, most successful demonstrations of molecular logic gates have been performed in free solution[4–6], but there are advantages to be had from spreading the active logical molecules on a surface. For example, smaller quantities of material are needed and the logic gates are easier to recover and reuse; more bang for our buck.

an inactive form of haemoglobin, the other famed member of the globin family.

Surprisingly low concentrations of nitrite, near the physiological level, change blood-flow properties and may modulate blood pressure[9]. Low levels of nitrite also act to suppress the activity of mitochondria, the organelles that are the main source of energy generation in a cell. In so doing, nitrite reduces the generation of mitochondrion-derived damaging reactive oxygen species (ROS)[2] in tissues suffering from ischaemia and from reperfusion damage. These conditions, respectively, are an interruption of blood flow and oxygen delivery to tissues caused by blocked blood vessels, and damage caused by a surge in the production of ROS.

Understanding how nitrite exerts its beneficial effects is of obvious physiological and medical significance[10]. Those effects are mainly linked to its reductive conversion to nitric oxide (NO), which relaxes the smooth-muscle lining of arterioles, the intermediate blood vessels between arteries and capillaries, and so dilates them[11]. But nitric oxide also acts on heart muscle, where it decreases muscle contractility and heart-beat rate by binding to mitochondrial proteins and suppressing the electron-transport chain that is essential in energy production[12]. This reduces the demand for oxygen and the ROS-induced damage during times of local oxygen deficiency. Nitric oxide synthases are generally regarded as the source of arteriolar nitric oxide by metabolizing the amino acid L-arginine. But there is increasing evidence that tissue nitric oxide can also be generated from nitrite through a nitrite reductase activity mediated by proteins such as xanthine oxidoreductase, deoxyhaemoglobin and deoxymyoglobin, or by a non-enzymatic reaction under acidic conditions[13].

The hypothesis that haemoglobin and myoglobin can mediate production of nitric oxide has been particularly controversial, because these molecules had been regarded as dioxygenases whose principal function is to scavenge surplus nitric oxide[4]. Hendgen-Cotta *et al.*[6], however, now clarify matters (Fig. 1). With myoglobin-knockout mice, they provide evidence that nitrite is the source of nitric oxide that protects heart cells following an episode of ischaemia or reperfusion, and — crucially — that deoxymyoglobin acts as the necessary nitrite reductase.

Thus, mice without myoglobin cannot generate nitrite-dependent nitric oxide, or nitrosylate heart muscle, or generate the cyclic GMP that is the essential signalling agent in producing vasodilation. Furthermore, compared with normal mice, the hearts of the knockout mice do not recover from experimentally imposed ischaemia; and these mice show no evidence of nitrite-induced reduction in the damage to heart tissue caused by blood-vessel blockage. On this evidence, then, myoglobin is both a reductase and a dioxygenase — it respectively produces and scavenges nitric oxide in deoxygenated and oxygenated conditions.

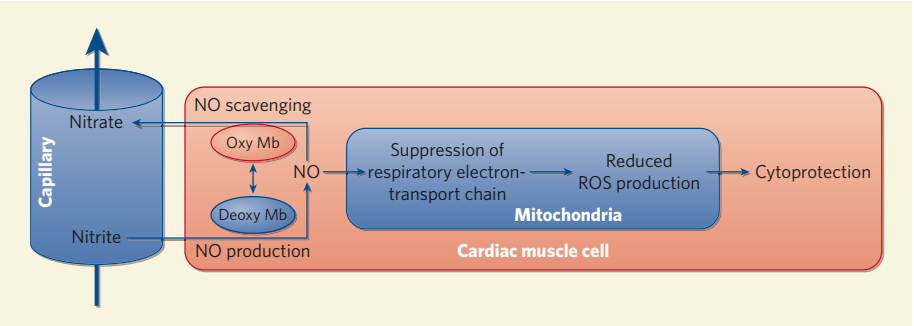


Figure 1 | Myoglobin as a reductase and an oxygenase. Myoglobin (Mb) respectively produces and scavenges nitric oxide (NO) in deoxygenated and oxygenated conditions. In its action as a reductase[6], deoxygenated myoglobin generates NO from circulating nitrite. This process is activated in cardiac muscle cells under hypoxic stress, where it suppresses the production of damaging reactive oxygen species (ROS) in mitochondria, so protecting the muscle cells from damage. Excess NO is reconverted to nitrate by oxymyoglobin acting as a dioxygenase.

The haem-containing globin superfamily includes two of the most-studied proteins in history, so it is astonishing to find that one of them has hitherto unidentified functions. But we can be confident that there will be more surprises, for at least two reasons. Myoglobin has been identified in tissues such as brain and liver[14], suggesting that its versatility in function applies to more than just muscle. And no role has yet been established for some other members of the superfamily, such as cytoglobin and neuroglobin[15]. ■

Andrew Cossins and Michael Berenbrink are in the School of Biological Sciences, University of Liverpool, Liverpool L69 7ZB, UK.
e-mail: cossins@liverpool.ac.uk

1. Garry, D. J. *et al. Nature* **395,** 905–908 (1998).
2. Shiva, S. *et al. Circ. Res.* **100,** 654–661 (2007).
3. Rayner, B. S., Wu, B.-J., Raftery, M., Stocker, R. & Witting, P. K. *J. Biol. Chem.* **280,** 9985–9993 (2005).
4. Brunori, M. *Trends Biochem. Sci.* **26,** 209–210 (2001).
5. Flögel, U. *et al. Proc. Natl Acad. Sci. USA* **98,** 735–740 (2001).
6. Hendgen-Cotta, U. B. *et al. Proc. Natl Acad. Sci. USA* doi:10.1073/pnas.0801336105 (2008).
7. Jensen, F. B. *Comp. Biochem. Physiol. Mol. Integr. Physiol.* **135,** 9–24 (2003).
8. Powlson, D. S. *et al. J. Environ. Qual.* **37,** 291–295 (2008).
9. Cosby, K. *et al. Nature Med.* **9,** 1498–1505 (2003).
10. Kumar, D. *et al. Proc. Natl Acad. Sci. USA* **105,** 7540–7545 (2008).
11. Furchgott, R. F. *Angew. Chem. Int. Edn* **38,** 1870–1880 (1999).
12. Rassaf, T. *et al. Circ. Res.* **100,** 1749–1754 (2007).
13. Lundberg, J. O., Weitzberg, E. & Gladwin, M. T. *Nature Rev. Drug Discov.* **7,** 156–167 (2008).
14. Fraser, J. *et al. Proc. Natl Acad. Sci. USA* **103,** 2977–2981 (2006).
15. Hankeln, T. *et al. J. Inorg. Biochem.* **99,** 110–119 (2005).

MOLECULAR COMPUTING

A layer of logic

A. Prasanna de Silva

Silicon chips have thousands of electronic logic gates etched on them. But there are other ways to decorate monolithic surfaces with logic gates, as a system using metal complexes secured to glass slides shows.

Logic lies at the heart of modern computers[1], and the components that carry out its operations are logic gates. On the basis of a number of digital inputs (the presence or absence of some condition conventionally represented as 1's or 0's), a logic gate produces a single digital output. For example, a simple AND gate produces a 1 output when both of its inputs are 1, and a 0 output if either or both inputs are 0. In electronic computers, logic gates are sculpted on the surface of silicon wafers. Their inputs and outputs are electrical voltages, but these are not the only systems that can form logic circuits. Writing in *Angewandte Chemie*, Gupta and van der Boom[2] describe a series of redox-active molecules anchored on glass surfaces that function as chemical logic gates.

The practice of using molecules as logic gates is about 15 years old[3], and could be said to be entering a turbulent adolescence. Although electronic logic circuits have been highly successful, molecular logic gates have some potential advantages — not least that they can operate in much smaller spaces. Molecular logic gates use concentrations of specific chemicals as their inputs, and the production of a particular chemical species, often accompanied by a colour change or similarly easy-to-identify property, as their outputs. So far, most successful demonstrations of molecular logic gates have been performed in free solution[4–6], but there are advantages to be had from spreading the active logical molecules on a surface. For example, smaller quantities of material are needed and the logic gates are easier to recover and reuse; more bang for our buck.

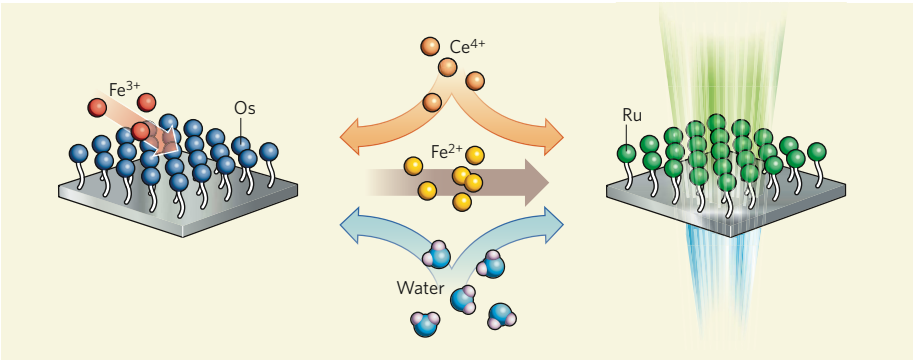


Figure 1 | Computing on glass. Gupta and van der Boom[2] tethered organic complexes containing osmium (Os) or ruthenium (Ru) to separate glass surfaces to form molecular logic gates. The osmium-containing gate is an OR gate in which the positive inputs are the presence of Fe^{3+} and water; Fe^{2+} is the positive output. The Fe^{2+} reduces the ruthenium-containing complexes, changing their absorption of 463-nm-wavelength light, which is used as the output of the gate. However, Ce^{4+} prevents this change, making the ruthenium-containing complexes AND gates for the presence of Fe^{2+} and absence of Ce^{4+}. Bathing glass slides containing these two complexes in the same solution effectively connects the molecular logic gates in series, creating a compound device with three inputs: Fe^{3+}, Ce^{4+} and water, and a single, spectroscopic output.

For their logic gates, Gupta and van der Boom[2] used organic molecules containing the metals osmium (Os) or ruthenium (Ru), which they tethered to a glass surface to form layers one molecule thick. The metals bound in the organic compounds can exist in one of two oxidation states, and this state provides the output for the logic gate. Applying the correct oxidizing or reducing agent to the immobilized complexes under suitable conditions converts the metal from one oxidation state to the other. These oxidation states have different colours. Complexes containing the reduced form of osmium, Os^{2+}, absorb light of wavelengths around 516 nm more strongly than the oxidized form, Os^{3+}, and less strongly at wavelengths around 317 nm. Measuring the absorption of light at judiciously chosen wavelengths identifies the oxidation state of the logic gate and so provides the output for the device.

Specific redox agents and solvents at high or low concentrations are used as inputs to the devices. For example, in the presence of silver ions and the solvent dichloromethane, the osmium-containing molecules will pass into the oxidized form and so strongly absorb light of wavelength 317 nm. In the absence of silver, or if a different solvent (for example acetone) is used, then osmium will remain in the initial, reduced state and so absorb weakly at 317 nm. In mathematical terms, this is a two-input AND gate: the presence or absence of silver and dichloromethane are the two binary inputs (1 for presence, 0 for absence) and the absorption at 317 nm is the binary output (1 for high absorption, 0 for low absorption).

By using combinations of redox agents, solvents and various metal compounds in different oxidation states, and monitoring the molecular logic element at different wavelengths, Gupta and van der Boom have produced an extensive set of two- and three-input logic gates[2]. Some of the three-input cases are equivalent to circuits involving several elementary logic gates with their inputs and outputs linked. One example using water, a nitric oxide derivative called nitrosonium ions (NO^+) and cerium ions (Ce^{4+}) as inputs requires a diagram involving three NOT gates and two AND gates. The integration of logic elements in these cases is not physical because there is no wiring. This is functional logic integration.

The same authors have shown previously that the chemical product arising from a redox reaction involving one glass-bound metal compound can serve as a reactant with a second glass-bound compound[7]. They now use this idea to connect two different molecular logic gates (Fig. 1). The reduced form of the osmium-bearing compound will reduce Fe^{3+} to produce Fe^{2+}. Gupta and van der Boom arranged things so that any Fe^{2+} thus produced could diffuse across to a ruthenium-bearing compound and reduce it to the Ru^{2+} state, but only if Ce^{4+} is absent; a variation on a logical AND gate. The relevant output of the osmium-containing device is now not a change in light absorption but the presence or absence of reduced iron. The oxidation state of iron is then used as an input to the ruthenium device, the oxidation state — and so output — of which is monitored by changes in the absorption of light of wavelength 463 nm. The osmium and ruthenium compounds do not directly interact or interfere with each other because they are immobilized and so cannot come into contact. This approach serially integrates separate molecular logic devices in a physical sense, with the Fe^{2+} providing the wiring.

Molecular logic has undergone gradual development[4–6] during its first decade and a half, and the first practical applications are now coming on-stream[8]. With results such as those of Gupta and van der Boom[2], the field's coming of age cannot be too far away. ■

A. Prasanna de Silva is at the School of Chemistry and Chemical Engineering, Queen's University, Belfast BT9 5AG, UK.
e-mail: a.desilva@qub.ac.uk

1. Ben-Ari, M. *Mathematical Logic for Computer Science* (Prentice Hall, 1993).
2. Gupta, T. & van der Boom, M. E. *Angew. Chem. Int. Edn* **47,** 5322–5325 (2008).
3. de Silva, A. P., Gunaratne, N. H. Q. & McCoy, C. P. *Nature* **364,** 42–44 (1993).
4. Balzani, V., Credi, A. & Venturi, M. *Molecular Devices and Machines* 2nd edn (Wiley-VCH, 2008).
5. Pischel, U. *Angew. Chem. Int. Edn* **46,** 4026–4040 (2007).
6. de Silva, A. P. & Uchiyama, S. *Nature Nanotech.* **2,** 399–410 (2007).
7. Gupta, T. & van der Boom, M. E. *Angew. Chem. Int. Edn* **47,** 2260–2262 (2008).
8. de Silva, A. P., James, M. R., McKinney, B. O. F., Pears, D. A. & Weir, S. M. *Nature Mater.* **5,** 787–790 (2006).

ALZHEIMER'S DISEASE

Moving towards a vaccine

David M. Holtzman

An agent that clears disease-associated amyloid aggregates from the brains of patients with Alzheimer's disease does not alleviate disease progression. Yet this disappointing news should not rule out such potential therapies.

Alzheimer's disease is the most common cause of dementia. But although there are some drugs that can slightly alleviate its symptoms, there is currently no treatment that can prevent this neurodegenerative disorder, delay its onset or slow its progress. Genetic, biochemical and animal studies have provided strong evidence that brain accumulation of a peptide of 38–43 amino acids, known as amyloid-β (Aβ), in the form of plaques and small aggregates (oligomers) is essential for the development of Alzheimer's disease[1]. So, will a vaccine that removes Aβ aggregates from a patient's brain delay or even cure this disease? Writing in *The Lancet*, Holmes *et al.*[2] report the findings of a preliminary study that aimed to answer this question.

The search for a treatment for Alzheimer's disease has led scientists to investigate ways of decreasing the accumulation of Aβ in the brain by reducing Aβ production, inhibiting its aggregation or enhancing its clearance. Previous work[3] on mice genetically modified to accumulate Aβ in their brains showed that, when these animals are immunized with external Aβ, they develop anti-Aβ antibodies,

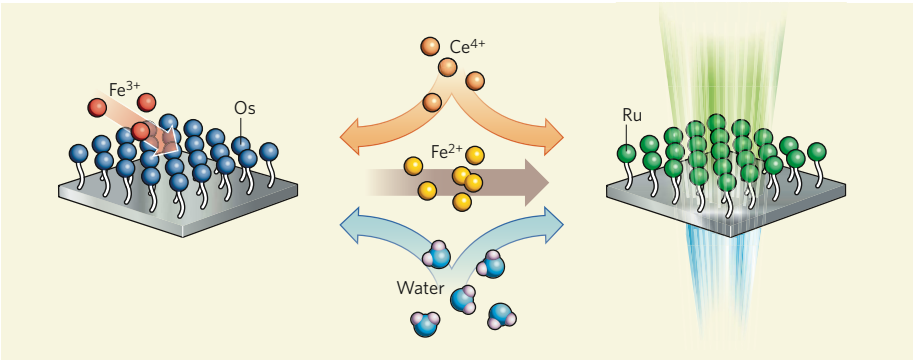


Figure 1 | Computing on glass. Gupta and van der Boom[2] tethered organic complexes containing osmium (Os) or ruthenium (Ru) to separate glass surfaces to form molecular logic gates. The osmium-containing gate is an OR gate in which the positive inputs are the presence of Fe^{3+} and water; Fe^{2+} is the positive output. The Fe^{2+} reduces the ruthenium-containing complexes, changing their absorption of 463-nm-wavelength light, which is used as the output of the gate. However, Ce^{4+} prevents this change, making the ruthenium-containing complexes AND gates for the presence of Fe^{2+} and absence of Ce^{4+}. Bathing glass slides containing these two complexes in the same solution effectively connects the molecular logic gates in series, creating a compound device with three inputs: Fe^{3+}, Ce^{4+} and water, and a single, spectroscopic output.

For their logic gates, Gupta and van der Boom[2] used organic molecules containing the metals osmium (Os) or ruthenium (Ru), which they tethered to a glass surface to form layers one molecule thick. The metals bound in the organic compounds can exist in one of two oxidation states, and this state provides the output for the logic gate. Applying the correct oxidizing or reducing agent to the immobilized complexes under suitable conditions converts the metal from one oxidation state to the other. These oxidation states have different colours. Complexes containing the reduced form of osmium, Os^{2+}, absorb light of wavelengths around 516 nm more strongly than the oxidized form, Os^{3+}, and less strongly at wavelengths around 317 nm. Measuring the absorption of light at judiciously chosen wavelengths identifies the oxidation state of the logic gate and so provides the output for the device.

Specific redox agents and solvents at high or low concentrations are used as inputs to the devices. For example, in the presence of silver ions and the solvent dichloromethane, the osmium-containing molecules will pass into the oxidized form and so strongly absorb light of wavelength 317 nm. In the absence of silver, or if a different solvent (for example acetone) is used, then osmium will remain in the initial, reduced state and so absorb weakly at 317 nm. In mathematical terms, this is a two-input AND gate: the presence or absence of silver and dichloromethane are the two binary inputs (1 for presence, 0 for absence) and the absorption at 317 nm is the binary output (1 for high absorption, 0 for low absorption).

By using combinations of redox agents, solvents and various metal compounds in different oxidation states, and monitoring the molecular logic element at different wavelengths, Gupta and van der Boom have produced an extensive set of two- and three-input logic gates[2]. Some of the three-input cases are equivalent to circuits involving several elementary logic gates with their inputs and outputs linked. One example using water, a nitric oxide derivative called nitrosonium ions (NO^+) and cerium ions (Ce^{4+}) as inputs requires a diagram involving three NOT gates and two AND gates. The integration of logic elements in these cases is not physical because there is no wiring. This is functional logic integration.

The same authors have shown previously that the chemical product arising from a redox reaction involving one glass-bound metal compound can serve as a reactant with a second glass-bound compound[7]. They now use this idea to connect two different molecular logic gates (Fig. 1). The reduced form of the osmium-bearing compound will reduce Fe^{3+} to produce Fe^{2+}. Gupta and van der Boom arranged things so that any Fe^{2+} thus produced could diffuse across to a ruthenium-bearing compound and reduce it to the Ru^{2+} state, but only if Ce^{4+} is absent; a variation on a logical AND gate. The relevant output of the osmium-containing device is now not a change in light absorption but the presence or absence of reduced iron. The oxidation state of iron is then used as an input to the ruthenium device, the oxidation state — and so output — of which is monitored by changes in the absorption of light of wavelength 463 nm. The osmium and ruthenium compounds do not directly interact or interfere with each other because they are immobilized and so cannot come into contact. This approach serially integrates separate molecular logic devices in a physical sense, with the Fe^{2+} providing the wiring.

Molecular logic has undergone gradual development[4–6] during its first decade and a half, and the first practical applications are now coming on-stream[8]. With results such as those of Gupta and van der Boom[2], the field's coming of age cannot be too far away. ■

A. Prasanna de Silva is at the School of Chemistry and Chemical Engineering, Queen's University, Belfast BT9 5AG, UK.
e-mail: a.desilva@qub.ac.uk

1. Ben-Ari, M. *Mathematical Logic for Computer Science* (Prentice Hall, 1993).
2. Gupta, T. & van der Boom, M. E. *Angew. Chem. Int. Edn* **47,** 5322–5325 (2008).
3. de Silva, A. P., Gunaratne, N. H. Q. & McCoy, C. P. *Nature* **364,** 42–44 (1993).
4. Balzani, V., Credi, A. & Venturi, M. *Molecular Devices and Machines* 2nd edn (Wiley-VCH, 2008).
5. Pischel, U. *Angew. Chem. Int. Edn* **46,** 4026–4040 (2007).
6. de Silva, A. P. & Uchiyama, S. *Nature Nanotech.* **2,** 399–410 (2007).
7. Gupta, T. & van der Boom, M. E. *Angew. Chem. Int. Edn* **47,** 2260–2262 (2008).
8. de Silva, A. P., James, M. R., McKinney, B. O. F., Pears, D. A. & Weir, S. M. *Nature Mater.* **5,** 787–790 (2006).

ALZHEIMER'S DISEASE

Moving towards a vaccine

David M. Holtzman

An agent that clears disease-associated amyloid aggregates from the brains of patients with Alzheimer's disease does not alleviate disease progression. Yet this disappointing news should not rule out such potential therapies.

Alzheimer's disease is the most common cause of dementia. But although there are some drugs that can slightly alleviate its symptoms, there is currently no treatment that can prevent this neurodegenerative disorder, delay its onset or slow its progress. Genetic, biochemical and animal studies have provided strong evidence that brain accumulation of a peptide of 38–43 amino acids, known as amyloid-β (Aβ), in the form of plaques and small aggregates (oligomers) is essential for the development of Alzheimer's disease[1]. So, will a vaccine that removes Aβ aggregates from a patient's brain delay or even cure this disease? Writing in *The Lancet*, Holmes *et al.*[2] report the findings of a preliminary study that aimed to answer this question.

The search for a treatment for Alzheimer's disease has led scientists to investigate ways of decreasing the accumulation of Aβ in the brain by reducing Aβ production, inhibiting its aggregation or enhancing its clearance. Previous work[3] on mice genetically modified to accumulate Aβ in their brains showed that, when these animals are immunized with external Aβ, they develop anti-Aβ antibodies,

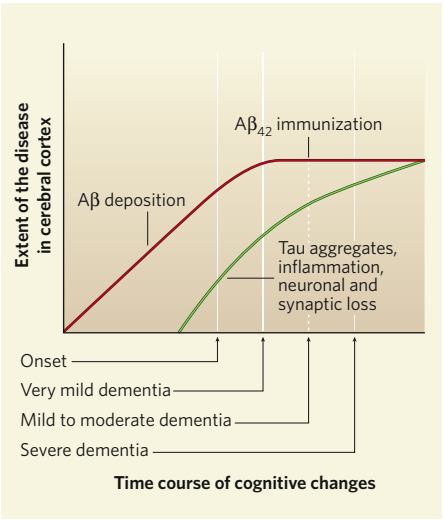


Figure 1 | Aβ deposition, immunization and cognitive change in Alzheimer's disease. Holmes *et al.*[2] observed no clinical benefits from actively immunizing Alzheimer's patients with $A\beta_{42}$. This could be because the immunizations started too late in the course of the disease — that is, after the patients had developed mild to moderate dementia. Brain deposition of the Aβ protein begins roughly 10–15 years before the onset of the cognitive decline that is recognized as dementia due to Alzheimer's disease. It is likely that Aβ accumulation peaks at the stage of very mild dementia, before there is large-scale brain deposition of another disease-associated protein, tau. Inflammatory changes and neuronal and synaptic loss probably also begin several years before the onset of cognitive decline.

and brain amyloid-plaque formation decreases. This promising observation suggested that such an immunization procedure might prevent and even reverse the disease in humans too.

In subsequent studies, animals were either actively immunized with Aβ plus various adjuvants (complementary agents that boost the effect of a treatment), or passively immunized through the direct introduction of external, specific anti-Aβ antibodies. These treatments decreased Aβ accumulation and plaque formation, as well as improving the animals' cognitive performance[4]. The studies formed the basis of the first human phase I clinical trials[5] in 2000 in which 80 patients with Alzheimer's disease were immunized with the 42-amino-acid Aβ peptide called $A\beta_{42}$ (AN1792; Elan Pharmaceuticals). Although no complications were associated with this trial, a subsequent, larger phase II trial[6] was halted after 6% of the subjects developed meningoencephalitis — inflammation of the brain and the membranes that cover it.

Holmes *et al.*[2] now report on the six-year follow-up of the original 80 subjects from the phase I $A\beta_{42}$ immunization trial. Eight patients who were examined post-mortem showed evidence of a significant reduction in brain amyloid plaques, together with increased blood anti-Aβ antibody levels. Despite plaque reduction, however, the authors found no evidence of a delay in progression to either severe dementia or death in these patients when compared with control subjects immunized with adjuvants only.

There could be several reasons for this paradoxical observation. First, the study was not designed to detect small but potentially significant changes in disease progress in the immunized group. Moreover, when the normally monomeric, soluble Aβ accumulates in the brain, it seems to build up in two forms: insoluble amyloid plaques and the smaller oligomeric forms that may be particularly toxic. Holmes *et al.* show that immunization removes plaques, but whether there is a clearance or neutralization of oligomers is not known. Nonetheless, two patients with virtually complete amyloid-plaque clearance as a presumed consequence of immunization still progressed from mild/moderate dementia to severe dementia.

The absence of any long-term clinical benefit following immunization with $A\beta_{42}$ could very well also be due to the fact that the treatment started too late in the disease process. The large-scale aggregation and accumulation of Aβ in the neocortex of the human brain seems to begin some 10–15 years before the onset of any symptoms or signs of dementia due to the disease[7]. In fact, Aβ accumulation has probably peaked by the time individuals with the disease have very mild dementia[7,8] (Fig. 1). In Holmes and colleagues' study[2], Aβ immunization started when subjects already had mild to moderate dementia.

Furthermore, Alzheimer's disease, like other neurodegenerative diseases, is a disorder of protein misfolding. In addition to Aβ accumulation, a normally soluble protein called tau becomes insoluble and accumulates in the brain. Although in some brain regions tau accumulates at the same time and even earlier than Aβ, its more extensive build-up in the neocortex seems to occur somewhat later than Aβ accumulation, and continues right through the clinical course of the disease (Fig. 1). More importantly, by the time patients have even very mild dementia, there is already significant neuronal death in certain brain regions that subserve memory[8]. Finally, animal studies suggest that, although Aβ accumulation drives the formation of abnormal tau aggregates[9], once this latter process is at an advanced stage, neutralization or removal of Aβ does not seem to reverse the later-stage tau-associated pathology[10].

So is there still reason to think that Aβ is a good therapeutic target for Alzheimer's disease? Definitely. But it must be remembered that both human and animal studies indicate that, for the greatest chance of success, Aβ must be targeted during the preclinical or very early clinical stages of the disease. Preclinical stages of Alzheimer's disease seem to be detectable through imaging and using biomarkers[11]. So it should be possible to design 'prevention' trials with smaller patient groups at earlier stages of the disease in the hope of having a greater

50 YEARS AGO

As from June 1, 1958, the Clean Air Act, 1956 ... is fully in force, and it will be an offence punishable by fine to emit 'dark smoke' (defined as as dark as or darker than shade 2 on the Ringelmann Chart) from any chimney in England and Wales. The ban on dark smoke applies to all buildings, and to railway engines and ships, but it will chiefly affect industrial and commercial premises.

ALSO:

It was announced in the April issues of the *Physical Review* that beginning with the issue dated July 1 the *Physical Review* will no longer carry the feature 'Letters to the Editor', but that the 'Letters' will be published separately in a supplementary fortnightly journal to be called *Physical Review Letters*. The new journal ... will initially be sent free of charge to all subscribers of the *Physical Review*, though from January 1959 onwards a subscription will be charged.

From *Nature* 26 July 1958.

100 YEARS AGO

A noteworthy paper upon "The Limit of School Children's Capacity for Attention" was read by Prof. W. Phillips. After referring to the various experimental inquiries into this question, which have involved the use of various forms of Mosso's ergograph, or fatigue recorder, and Griesbach's æsthesiometer, and many experiments designed to test the rate of deterioration in mental work done at different times of the day and on different days of the week, Prof. Phillips discussed the useful results which all this work has led to: ... (1) during an ordinary school session children can maintain a more even degree of attention, if one or two intervals of rest are included ... (2) a child's attention wanes more rapidly in the afternoon than in the morning ... (3) The various branches of mathematics seem, *ceteris paribus*, to make a greater demand on the attention than most other subjects.

From *Nature* 23 July 1908.

50 & 100 YEARS AGO

impact on this devastating condition. Finally, treatments that would target not just Aβ but also other disease pathways, such as tau accumulation and inflammation, might form the ideal approach. ■

David M. Holtzman is in the Department of Neurology and Developmental Biology, Washington University, 660 South Euclid Avenue, St Louis, Missouri 63110, USA.
e-mail: holtzman@neuro.wustl.edu

1. Hardy, J. & Selkoe, D. J. *Science* **297,** 353–356 (2002).
2. Holmes, C. *et al. Lancet* **372,** 216–223 (2008).
3. Schenk, D. *et al. Nature* **400,** 173–177 (1999).
4. Brody, D. L. & Holtzman, D. M. *Annu. Rev. Neurosci.* **31,** 175–193 (2008).
5. Bayer, A. J. *et al. Neurology* **64,** 94–101 (2005).
6. Orgogozo, J.-M. *et al. Neurology* **61,** 46–54 (2003).
7. Price, J. L. & Morris, J. C. *Ann. Neurol.* **45,** 358–368 (1999).
8. Gómez-Isla, T. *et al. J. Neurosci.* **16,** 4491–4500 (1996).
9. Lewis, J. *et al. Science* **293,** 1487–1491 (2001).
10. Oddo, S. *et al. Neuron* **43,** 321–332 (2004).
11. Fagan, A. M. *et al. Ann. Neurol.* **59,** 512–519 (2006).

MATERIALS SCIENCE

A tale of two tilings

Sharon C. Glotzer and Aaron S. Keys

What do you get when you cross a crystal with a quasicrystal? The answer is a structure that links the ancient tiles of Archimedes, the iconic Fibonacci sequence of numbers and a book from the seventeenth century.

Quasicrystals are mosaic-like arrangements of atoms that have symmetries once thought to be impossible for crystals to adopt[1]. Primarily observed in certain metal alloys, these unusual structures are stronger and less deformable than analogous regular crystals, and have unusual frictional, catalytic and optical properties. Several applications have been proposed for quasicrystals — for example, some could be used as materials for photonic circuits[2]. But for this application to be realized, the atomic dimensions of a quasicrystal must first be scaled up almost 1,000-fold. On page 501 of this issue[3], Mikhael *et al.* describe quasicrystals at just such a scale, made from microscopic plastic beads. To their surprise, they also discovered a new kind of structure: a rare type of one-dimensional quasicrystal that can be thought of as a cross between a two-dimensional quasicrystal and a regular crystal.

Mikhael *et al.* grow single layers of colloidal particles on a templated surface designed to attract those particles and arrange them into pentagons — the primary motif of a quasicrystal with tenfold (decagonal) symmetry. They do this by arranging five laser beams to form an interference pattern that confers decagonal symmetry to the surface's potential, which interacts with the particles[4]. By tuning the strength of the surface potential using the lasers, the team controls the formation of the growing structures: regular crystals form when particle–particle interactions dominate, and quasicrystals form when particle–surface interactions dominate. The resulting quasicrystals exhibit rings of ten particles surrounding a central particle (see Fig. 1c on page 501).

Quasicrystals are often considered to be intermediate between glasses (amorphous solids) and crystals[5]. But can a structure be intermediate between a crystal and a quasicrystal? Conventional thinking says no — long-range ordering must be either periodic (crystalline) or aperiodic (quasicrystalline), with little room in between. But Mikhael *et al.*[3] find that, when the particle–particle and particle–surface interactions in their system are similar in strength, an intermediate phase forms that combines elements of both crystalline and quasicrystalline ordering. In fact, the particles assemble into something that closely resembles an Archimedean tiling pattern.

Archimedean tilings are periodic arrangements of regular polygons laid edge-to-edge in a plane. Their defining feature is that only one kind of vertex must exist — that is, where the corners of the polygons meet at a point, any given corner must always meet the same combination of corners from other polygons. Archimedean tilings have been used in art and architecture since antiquity, but it was the astronomer Johannes Kepler who first classified them in his book, *Harmonices Mundi*, in 1619. Kepler showed that there are eleven different kinds of tiling, eight of which contain more than one type of regular polygon. One tiling consists entirely of equilateral triangles, and is denoted (3^6) to indicate that six triangles meet at each vertex. This structure describes the crystal that Mikhael *et al.* observe when particle–particle interactions dominate. Another Archimedean tiling denoted $(3^3,4^2)$ consists of alternating rows of squares and triangles.

Mikhael and colleagues' new arrangement of particles is similar to the $(3^3,4^2)$ arrangement, with some (3^6) vertex configurations added in a peculiar way. The particles form alternating rows of squares and triangles, which are interrupted intermittently by 'defects' — additional rows of triangles that introduce (3^6) vertex configurations to the tiling (Fig. 1). The particles still align locally with the decagonal, quasicrystalline template, but a mismatch between the periodic tiling and the aperiodic substrate arises over longer distances. This is where the defects come in — the extra rows of triangles correct the mismatches.

The defects result in two distinct 'unit cells' (basic arrangements from which the tilings are constructed) that have different heights (Fig. 1). The heights of the cells correspond to the heights of the large and small pentagonal arrangements that are conferred on the particles by the underlying template field. The cells stack in a quasiperiodic pattern known as a Fibonacci chain. Named after a famous mathematician of the Middle Ages, this pattern is often found in nature, and describes the

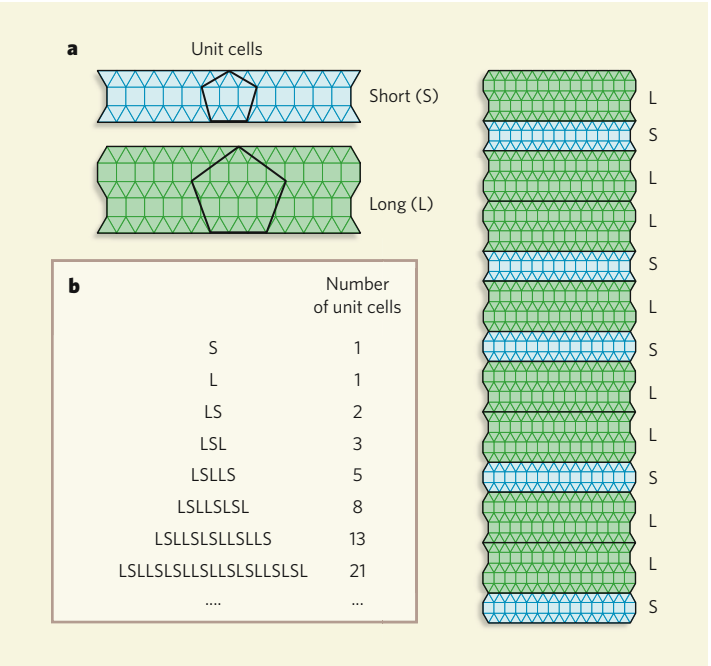


Figure 1 | A tiling structure based on the Fibonacci chain. a, Mikhael *et al.*[1] have discovered a new arrangement that can be adopted by particles in two dimensions. The structure, shown here in idealized form, consists of two unit cells of different widths (short, S, or long, L) that stack up on top of each other. The heights of the cells correspond to the heights of the small and large pentagons, whose size ratio is given by the 'golden mean'. Particles sit at the vertices of the tiles. **b,** The order of unit cells is described by a Fibonacci chain — a quasiperiodic sequence that starts from just one unit cell and expands by applying the substitution rules L→LS, S→L at each step. The sequence with 13 elements describes the arrangement of unit cells in the structure shown on the right.

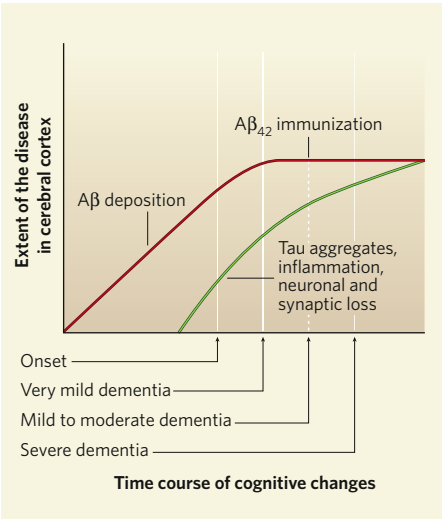


Figure 1 | Aβ deposition, immunization and cognitive change in Alzheimer's disease. Holmes *et al.*[2] observed no clinical benefits from actively immunizing Alzheimer's patients with $A\beta_{42}$. This could be because the immunizations started too late in the course of the disease — that is, after the patients had developed mild to moderate dementia. Brain deposition of the Aβ protein begins roughly 10–15 years before the onset of the cognitive decline that is recognized as dementia due to Alzheimer's disease. It is likely that Aβ accumulation peaks at the stage of very mild dementia, before there is large-scale brain deposition of another disease-associated protein, tau. Inflammatory changes and neuronal and synaptic loss probably also begin several years before the onset of cognitive decline.

and brain amyloid-plaque formation decreases. This promising observation suggested that such an immunization procedure might prevent and even reverse the disease in humans too.

In subsequent studies, animals were either actively immunized with Aβ plus various adjuvants (complementary agents that boost the effect of a treatment), or passively immunized through the direct introduction of external, specific anti-Aβ antibodies. These treatments decreased Aβ accumulation and plaque formation, as well as improving the animals' cognitive performance[4]. The studies formed the basis of the first human phase I clinical trials[5] in 2000 in which 80 patients with Alzheimer's disease were immunized with the 42-amino-acid Aβ peptide called $A\beta_{42}$ (AN1792; Elan Pharmaceuticals). Although no complications were associated with this trial, a subsequent, larger phase II trial[6] was halted after 6% of the subjects developed meningoencephalitis — inflammation of the brain and the membranes that cover it.

Holmes *et al.*[2] now report on the six-year follow-up of the original 80 subjects from the phase I $A\beta_{42}$ immunization trial. Eight patients who were examined post-mortem showed evidence of a significant reduction in brain amyloid plaques, together with increased blood anti-Aβ antibody levels. Despite plaque reduction, however, the authors found no evidence of a delay in progression to either severe dementia or death in these patients when compared with control subjects immunized with adjuvants only.

There could be several reasons for this paradoxical observation. First, the study was not designed to detect small but potentially significant changes in disease progress in the immunized group. Moreover, when the normally monomeric, soluble Aβ accumulates in the brain, it seems to build up in two forms: insoluble amyloid plaques and the smaller oligomeric forms that may be particularly toxic. Holmes *et al.* show that immunization removes plaques, but whether there is a clearance or neutralization of oligomers is not known. Nonetheless, two patients with virtually complete amyloid-plaque clearance as a presumed consequence of immunization still progressed from mild/moderate dementia to severe dementia.

The absence of any long-term clinical benefit following immunization with $A\beta_{42}$ could very well also be due to the fact that the treatment started too late in the disease process. The large-scale aggregation and accumulation of Aβ in the neocortex of the human brain seems to begin some 10–15 years before the onset of any symptoms or signs of dementia due to the disease[7]. In fact, Aβ accumulation has probably peaked by the time individuals with the disease have very mild dementia[7,8] (Fig. 1). In Holmes and colleagues' study[2], Aβ immunization started when subjects already had mild to moderate dementia.

Furthermore, Alzheimer's disease, like other neurodegenerative diseases, is a disorder of protein misfolding. In addition to Aβ accumulation, a normally soluble protein called tau becomes insoluble and accumulates in the brain. Although in some brain regions tau accumulates at the same time and even earlier than Aβ, its more extensive build-up in the neocortex seems to occur somewhat later than Aβ accumulation, and continues right through the clinical course of the disease (Fig. 1). More importantly, by the time patients have even very mild dementia, there is already significant neuronal death in certain brain regions that subserve memory[8]. Finally, animal studies suggest that, although Aβ accumulation drives the formation of abnormal tau aggregates[9], once this latter process is at an advanced stage, neutralization or removal of Aβ does not seem to reverse the later-stage tau-associated pathology[10].

So is there still reason to think that Aβ is a good therapeutic target for Alzheimer's disease? Definitely. But it must be remembered that both human and animal studies indicate that, for the greatest chance of success, Aβ must be targeted during the preclinical or very early clinical stages of the disease. Preclinical stages of Alzheimer's disease seem to be detectable through imaging and using biomarkers[11]. So it should be possible to design 'prevention' trials with smaller patient groups at earlier stages of the disease in the hope of having a greater

50 YEARS AGO

As from June 1, 1958, the Clean Air Act, 1956 ... is fully in force, and it will be an offence punishable by fine to emit 'dark smoke' (defined as as dark as or darker than shade 2 on the Ringelmann Chart) from any chimney in England and Wales. The ban on dark smoke applies to all buildings, and to railway engines and ships, but it will chiefly affect industrial and commercial premises.

ALSO:

It was announced in the April issues of the *Physical Review* that beginning with the issue dated July 1 the *Physical Review* will no longer carry the feature 'Letters to the Editor', but that the 'Letters' will be published separately in a supplementary fortnightly journal to be called *Physical Review Letters*. The new journal ... will initially be sent free of charge to all subscribers of the *Physical Review*, though from January 1959 onwards a subscription will be charged.

From *Nature* 26 July 1958.

100 YEARS AGO

A noteworthy paper upon "The Limit of School Children's Capacity for Attention" was read by Prof. W. Phillips. After referring to the various experimental inquiries into this question, which have involved the use of various forms of Mosso's ergograph, or fatigue recorder, and Griesbach's æsthesiometer, and many experiments designed to test the rate of deterioration in mental work done at different times of the day and on different days of the week, Prof. Phillips discussed the useful results which all this work has led to: ... (1) during an ordinary school session children can maintain a more even degree of attention, if one or two intervals of rest are included ... (2) a child's attention wanes more rapidly in the afternoon than in the morning ... (3) The various branches of mathematics seem, *ceteris paribus*, to make a greater demand on the attention than most other subjects.

From *Nature* 23 July 1908.

50 & 100 YEARS AGO

impact on this devastating condition. Finally, treatments that would target not just Aβ but also other disease pathways, such as tau accumulation and inflammation, might form the ideal approach. ■

David M. Holtzman is in the Department of Neurology and Developmental Biology, Washington University, 660 South Euclid Avenue, St Louis, Missouri 63110, USA.
e-mail: holtzman@neuro.wustl.edu

1. Hardy, J. & Selkoe, D. J. *Science* **297,** 353–356 (2002).
2. Holmes, C. *et al. Lancet* **372,** 216–223 (2008).
3. Schenk, D. *et al. Nature* **400,** 173–177 (1999).
4. Brody, D. L. & Holtzman, D. M. *Annu. Rev. Neurosci.* **31,** 175–193 (2008).
5. Bayer, A. J. *et al. Neurology* **64,** 94–101 (2005).
6. Orgogozo, J.-M. *et al. Neurology* **61,** 46–54 (2003).
7. Price, J. L. & Morris, J. C. *Ann. Neurol.* **45,** 358–368 (1999).
8. Gómez-Isla, T. *et al. J. Neurosci.* **16,** 4491–4500 (1996).
9. Lewis, J. *et al. Science* **293,** 1487–1491 (2001).
10. Oddo, S. *et al. Neuron* **43,** 321–332 (2004).
11. Fagan, A. M. *et al. Ann. Neurol.* **59,** 512–519 (2006).

MATERIALS SCIENCE

A tale of two tilings

Sharon C. Glotzer and Aaron S. Keys

What do you get when you cross a crystal with a quasicrystal? The answer is a structure that links the ancient tiles of Archimedes, the iconic Fibonacci sequence of numbers and a book from the seventeenth century.

Quasicrystals are mosaic-like arrangements of atoms that have symmetries once thought to be impossible for crystals to adopt[1]. Primarily observed in certain metal alloys, these unusual structures are stronger and less deformable than analogous regular crystals, and have unusual frictional, catalytic and optical properties. Several applications have been proposed for quasicrystals — for example, some could be used as materials for photonic circuits[2]. But for this application to be realized, the atomic dimensions of a quasicrystal must first be scaled up almost 1,000-fold. On page 501 of this issue[3], Mikhael *et al.* describe quasicrystals at just such a scale, made from microscopic plastic beads. To their surprise, they also discovered a new kind of structure: a rare type of one-dimensional quasicrystal that can be thought of as a cross between a two-dimensional quasicrystal and a regular crystal.

Mikhael *et al.* grow single layers of colloidal particles on a templated surface designed to attract those particles and arrange them into pentagons — the primary motif of a quasicrystal with tenfold (decagonal) symmetry. They do this by arranging five laser beams to form an interference pattern that confers decagonal symmetry to the surface's potential, which interacts with the particles[4]. By tuning the strength of the surface potential using the lasers, the team controls the formation of the growing structures: regular crystals form when particle–particle interactions dominate, and quasicrystals form when particle–surface interactions dominate. The resulting quasicrystals exhibit rings of ten particles surrounding a central particle (see Fig. 1c on page 501).

Quasicrystals are often considered to be intermediate between glasses (amorphous solids) and crystals[5]. But can a structure be intermediate between a crystal and a quasicrystal? Conventional thinking says no — long-range ordering must be either periodic (crystalline) or aperiodic (quasicrystalline), with little room in between. But Mikhael *et al.*[3] find that, when the particle–particle and particle–surface interactions in their system are similar in strength, an intermediate phase forms that combines elements of both crystalline and quasicrystalline ordering. In fact, the particles assemble into something that closely resembles an Archimedean tiling pattern.

Archimedean tilings are periodic arrangements of regular polygons laid edge-to-edge in a plane. Their defining feature is that only one kind of vertex must exist — that is, where the corners of the polygons meet at a point, any given corner must always meet the same combination of corners from other polygons. Archimedean tilings have been used in art and architecture since antiquity, but it was the astronomer Johannes Kepler who first classified them in his book, *Harmonices Mundi*, in 1619. Kepler showed that there are eleven different kinds of tiling, eight of which contain more than one type of regular polygon. One tiling consists entirely of equilateral triangles, and is denoted (3^6) to indicate that six triangles meet at each vertex. This structure describes the crystal that Mikhael *et al.* observe when particle–particle interactions dominate. Another Archimedean tiling denoted $(3^3,4^2)$ consists of alternating rows of squares and triangles.

Mikhael and colleagues' new arrangement of particles is similar to the $(3^3,4^2)$ arrangement, with some (3^6) vertex configurations added in a peculiar way. The particles form alternating rows of squares and triangles, which are interrupted intermittently by 'defects' — additional rows of triangles that introduce (3^6) vertex configurations to the tiling (Fig. 1). The particles still align locally with the decagonal, quasicrystalline template, but a mismatch between the periodic tiling and the aperiodic substrate arises over longer distances. This is where the defects come in — the extra rows of triangles correct the mismatches.

The defects result in two distinct 'unit cells' (basic arrangements from which the tilings are constructed) that have different heights (Fig. 1). The heights of the cells correspond to the heights of the large and small pentagonal arrangements that are conferred on the particles by the underlying template field. The cells stack in a quasiperiodic pattern known as a Fibonacci chain. Named after a famous mathematician of the Middle Ages, this pattern is often found in nature, and describes the

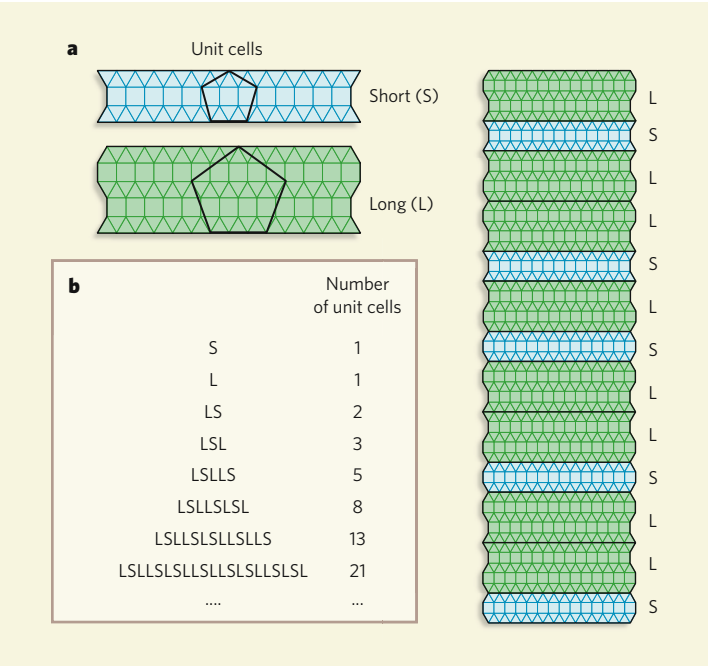


Figure 1 | A tiling structure based on the Fibonacci chain. a, Mikhael *et al.*[1] have discovered a new arrangement that can be adopted by particles in two dimensions. The structure, shown here in idealized form, consists of two unit cells of different widths (short, S, or long, L) that stack up on top of each other. The heights of the cells correspond to the heights of the small and large pentagons, whose size ratio is given by the 'golden mean'. Particles sit at the vertices of the tiles. **b,** The order of unit cells is described by a Fibonacci chain — a quasiperiodic sequence that starts from just one unit cell and expands by applying the substitution rules L→LS, S→L at each step. The sequence with 13 elements describes the arrangement of unit cells in the structure shown on the right.

structure of one-dimensional quasicrystals[1]. In Mikhael and colleagues' system, the Fibonacci chain determines the sequence of long and short cells. Because the Fibonacci chain is self-similar, the structure can also be described by simpler unit cells consisting of single rows of squares and triangles.

When grown on an icosahedral quasicrystalline surface, certain copper alloys also adopt a curious phase in which the atoms have a Fibonacci spacing[6]. The exact structure of the phase has not yet been identified, but its diffraction pattern is identical to that of Mikhael and colleagues' Archimedean-like arrangement of particles. If the two phases are indeed the same, it would demonstrate the universality of the underlying physics that controls the templated growth of these unusual structures. Furthermore, it would extend the growing use of colloids as minimal models of atoms for studying self-assembly[7] and other physical processes.

Archimedean tilings can also form from macromolecules that consist of three chemically distinct polymers, covalently bonded together at one end to form a three-armed 'star'[8]. Under certain conditions, these systems spontaneously form cylinders that have a cross-section corresponding to one of four Archimedean tilings. Two of these structures have useful optical properties and, like quasicrystals, hold promise for photonic applications.

It is not clear whether Archimedean-like tilings have a general role as intermediates between periodic and aperiodic structures. Such intermediates must be able to locally align with both the corresponding quasicrystal and crystal structures, and be able to incorporate aperiodically arranged defects. The ability to mix and match motifs may give Archimedean-tiling motifs a unique flexibility that makes them prone to forming aperiodic arrangements. For example, the dodecagonal quasicrystal[9], which exhibits 12-fold, rather than 10-fold, rotational symmetry, is made up of three different Archimedean vertex configurations, also called quasicrystal approximants.

Ultimately, we should not think of Mikhael and colleagues' structure[3] as a flawed Archimedean tiling. The underlying structure is a perfect Fibonacci chain, the elements of which are decorated with infinite rows of Archimedean tiles. From this perspective, it is a unique kind of one-dimensional quasicrystal, periodic in one dimension, but quasiperiodic in the other. This is what you get when you cross a crystal with a quasicrystal — a beguiling new tiling built upon iconic mathematical foundations. ■

Sharon C. Glotzer and Aaron S. Keys are in the Department of Chemical Engineering, University of Michigan, Ann Arbor, Michigan 48109-2136, USA.
e-mail: sglotzer@umich.edu

1. Janot, C. *Quasicrystals: A Primer* 2nd edn (Oxford Univ. Press, 1997).
2. Man, W., Megens, M., Steinhardt, P. J. & Chaikin, P. M. *Nature* **436,** 993–996 (2005).
3. Mikhael, J., Roth, J., Helden, L. & Bechinger, C. *Nature* **454,** 501–504 (2008).
4. Roichman, Y. & Grier, D. *Opt. Express* **13,** 5434–5439 (2005).
5. Steinhardt, P. J. *Nature* **452,** 43–44 (2008).
6. Ledieu, J. *et al. Phys. Rev. B* **72,** 035420 (2005).
7. Glotzer, S. C. & Solomon, M. J. *Nature Mater.* **6,** 557–562 (2007).
8. Ueda, K., Dotera, T. & Gemma, T. *Phys. Rev. B* **75,** 195122 (2007).
9. Keys, A. S. & Glotzer, S. C. *Phys. Rev. Lett.* **99,** 235503 (2007).

GENOMICS

Thoroughly modern meiosis

Michael Lichten

Meiotic recombination shuffles the genome, so each generation inherits a new combination of parental traits. Combining traditional and modern approaches, new work pinpoints where recombination occurs genome-wide.

During meiosis, a diploid cell (with two copies of each chromosome, one from each parent) undergoes two rounds of cell division, producing haploid gametes — in animals, these are sperm or eggs containing a single copy of each chromosome. Genetic recombination, which occurs at high levels during meiotic cell division, is crucial for chromosome separation in the diploid-to-haploid transition, and mixes parental genomic sequences to generate genetic diversity in the next generation. On page 479 of this issue, Mancera *et al.*[1] present the first comprehensive description of the meiotic recombination events that occur across an entire genome during a single meiosis, and provide tantalizing mechanistic insight into this process.

Much understanding of recombination mechanisms comes from studies in fungi such as budding yeast (*Saccharomyces cerevisiae*), where all four haploid meiotic segregants can be recovered. Genetic analysis of this ensemble of meiotic products, called a tetrad, led to the identification of fundamental features of meiotic recombination, such as gene conversion — the unidirectional replacement of genetic information on one parental chromosome by genetic information from another chromosome[2]. But such analysis is labour-intensive and limited in scope. Because of the limited availability of conventional genetic markers, only a small portion of the yeast genome has been examined in detail, and hundreds of tetrads need to be analysed to detect recombination events in sufficient numbers.

Mancera *et al.* overcame these limitations by combining traditional tetrad analysis with modern high-throughput molecular methods for the genome-wide scoring of sequence variations (polymorphisms). They mated two budding-yeast strains that are cross-fertile but have diverged evolutionarily, and that have sequence differences (mostly single-nucleotide changes) at almost 70,000 genomic sites[3]. Of these, 52,000 polymorphisms could be scored as genetic markers, allowing the detection of recombination throughout the genome at an unprecedented level of resolution and efficiency. The authors captured most of the recombination events that occurred in each of 51 separate meioses (6,289 events in total), and this allowed them to address several

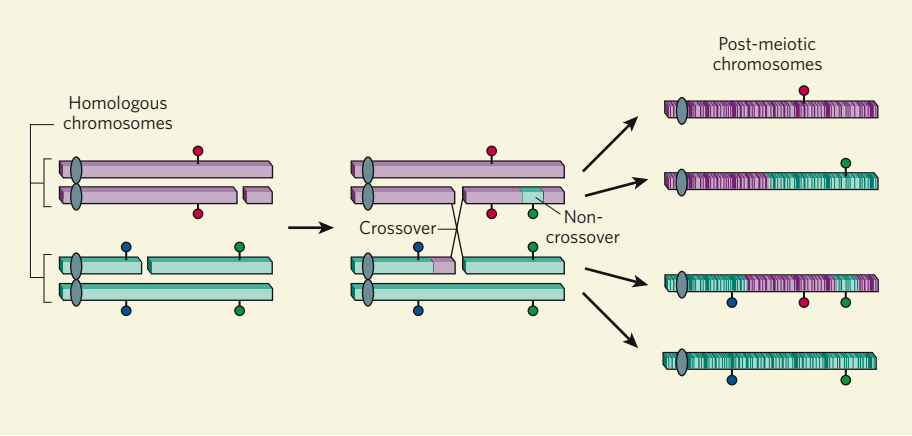


Figure 1 | Detecting meiotic recombination. Meiosis-induced DNA double-strand breaks are repaired by either crossover or non-crossover recombination, both of which are associated with gene conversion. Recombination between two parental homologous chromosomes can be detected only if they differ in genetic markers. In the example shown, tetrad analysis using conventional genetic markers (blue, red and green lollipops; centre) detects events with much less resolution than the high-density marker analysis (purple and green cross-hatches; right) used by Mancera *et al.*[1].

structure of one-dimensional quasicrystals[1]. In Mikhael and colleagues' system, the Fibonacci chain determines the sequence of long and short cells. Because the Fibonacci chain is self-similar, the structure can also be described by simpler unit cells consisting of single rows of squares and triangles.

When grown on an icosahedral quasicrystalline surface, certain copper alloys also adopt a curious phase in which the atoms have a Fibonacci spacing[6]. The exact structure of the phase has not yet been identified, but its diffraction pattern is identical to that of Mikhael and colleagues' Archimedean-like arrangement of particles. If the two phases are indeed the same, it would demonstrate the universality of the underlying physics that controls the templated growth of these unusual structures. Furthermore, it would extend the growing use of colloids as minimal models of atoms for studying self-assembly[7] and other physical processes.

Archimedean tilings can also form from macromolecules that consist of three chemically distinct polymers, covalently bonded together at one end to form a three-armed 'star'[8]. Under certain conditions, these systems spontaneously form cylinders that have a cross-section corresponding to one of four Archimedean tilings. Two of these structures have useful optical properties and, like quasicrystals, hold promise for photonic applications.

It is not clear whether Archimedean-like tilings have a general role as intermediates between periodic and aperiodic structures. Such intermediates must be able to locally align with both the corresponding quasicrystal and crystal structures, and be able to incorporate aperiodically arranged defects. The ability to mix and match motifs may give Archimedean-tiling motifs a unique flexibility that makes them prone to forming aperiodic arrangements. For example, the dodecagonal quasicrystal[9], which exhibits 12-fold, rather than 10-fold, rotational symmetry, is made up of three different Archimedean vertex configurations, also called quasicrystal approximants.

Ultimately, we should not think of Mikhael and colleagues' structure[3] as a flawed Archimedean tiling. The underlying structure is a perfect Fibonacci chain, the elements of which are decorated with infinite rows of Archimedean tiles. From this perspective, it is a unique kind of one-dimensional quasicrystal, periodic in one dimension, but quasiperiodic in the other. This is what you get when you cross a crystal with a quasicrystal — a beguiling new tiling built upon iconic mathematical foundations. ■

Sharon C. Glotzer and Aaron S. Keys are in the Department of Chemical Engineering, University of Michigan, Ann Arbor, Michigan 48109-2136, USA.
e-mail: sglotzer@umich.edu

1. Janot, C. *Quasicrystals: A Primer* 2nd edn (Oxford Univ. Press, 1997).
2. Man, W., Megens, M., Steinhardt, P. J. & Chaikin, P. M. *Nature* **436,** 993–996 (2005).
3. Mikhael, J., Roth, J., Helden, L. & Bechinger, C. *Nature* **454,** 501–504 (2008).
4. Roichman, Y. & Grier, D. *Opt. Express* **13,** 5434–5439 (2005).
5. Steinhardt, P. J. *Nature* **452,** 43–44 (2008).
6. Ledieu, J. *et al. Phys. Rev. B* **72,** 035420 (2005).
7. Glotzer, S. C. & Solomon, M. J. *Nature Mater.* **6,** 557–562 (2007).
8. Ueda, K., Dotera, T. & Gemma, T. *Phys. Rev. B* **75,** 195122 (2007).
9. Keys, A. S. & Glotzer, S. C. *Phys. Rev. Lett.* **99,** 235503 (2007).

GENOMICS

Thoroughly modern meiosis

Michael Lichten

Meiotic recombination shuffles the genome, so each generation inherits a new combination of parental traits. Combining traditional and modern approaches, new work pinpoints where recombination occurs genome-wide.

During meiosis, a diploid cell (with two copies of each chromosome, one from each parent) undergoes two rounds of cell division, producing haploid gametes — in animals, these are sperm or eggs containing a single copy of each chromosome. Genetic recombination, which occurs at high levels during meiotic cell division, is crucial for chromosome separation in the diploid-to-haploid transition, and mixes parental genomic sequences to generate genetic diversity in the next generation. On page 479 of this issue, Mancera *et al.*[1] present the first comprehensive description of the meiotic recombination events that occur across an entire genome during a single meiosis, and provide tantalizing mechanistic insight into this process.

Much understanding of recombination mechanisms comes from studies in fungi such as budding yeast (*Saccharomyces cerevisiae*), where all four haploid meiotic segregants can be recovered. Genetic analysis of this ensemble of meiotic products, called a tetrad, led to the identification of fundamental features of meiotic recombination, such as gene conversion — the unidirectional replacement of genetic information on one parental chromosome by genetic information from another chromosome[2]. But such analysis is labour-intensive and limited in scope. Because of the limited availability of conventional genetic markers, only a small portion of the yeast genome has been examined in detail, and hundreds of tetrads need to be analysed to detect recombination events in sufficient numbers.

Mancera *et al.* overcame these limitations by combining traditional tetrad analysis with modern high-throughput molecular methods for the genome-wide scoring of sequence variations (polymorphisms). They mated two budding-yeast strains that are cross-fertile but have diverged evolutionarily, and that have sequence differences (mostly single-nucleotide changes) at almost 70,000 genomic sites[3]. Of these, 52,000 polymorphisms could be scored as genetic markers, allowing the detection of recombination throughout the genome at an unprecedented level of resolution and efficiency. The authors captured most of the recombination events that occurred in each of 51 separate meioses (6,289 events in total), and this allowed them to address several

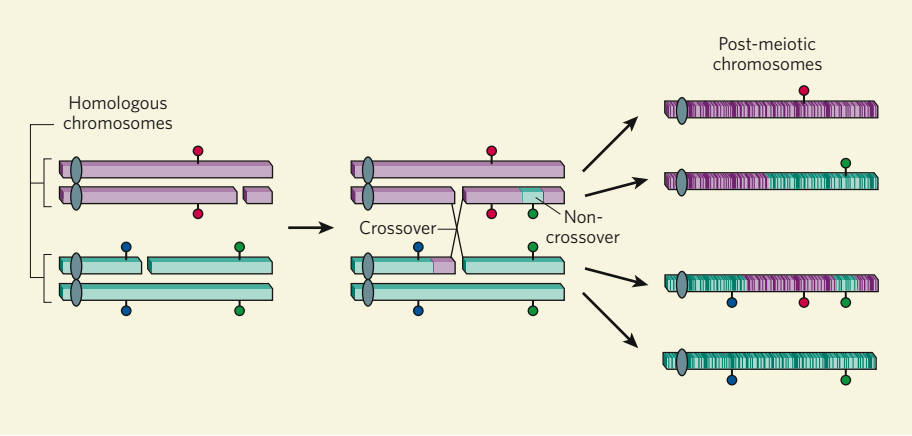


Figure 1 | Detecting meiotic recombination. Meiosis-induced DNA double-strand breaks are repaired by either crossover or non-crossover recombination, both of which are associated with gene conversion. Recombination between two parental homologous chromosomes can be detected only if they differ in genetic markers. In the example shown, tetrad analysis using conventional genetic markers (blue, red and green lollipops; centre) detects events with much less resolution than the high-density marker analysis (purple and green cross-hatches; right) used by Mancera *et al.*[1].

long-standing issues in meiotic recombination. I will focus on three: where recombination occurs; the type of recombination that occurs; and how recombination events are spaced along individual chromosomes.

Meiotic recombination is initiated by self-inflicted breaks in the double-stranded DNA; these breaks are then repaired by recombination with a homologous chromosome (Fig. 1). Genome-wide maps of meiotic double-strand breaks in budding yeast have indicated that most breaks occur in 'hotspots' that are distributed throughout the genome, with an average inter-hotspot distance of less than 10,000 bases[4]. Mancera and colleagues' recombination map[1] closely parallels these maps of double-strand breaks, with recombination occurring at almost all regions. In fact, the only truly recombination-cold regions of the yeast genome are associated with repeated sequences. These include mobile genetic elements, the DNA that encodes ribosomal RNA and sequences close to some chromosome ends.

The authors also investigated the relationship between the two outcomes of inter-homologue recombination: crossovers, where parental sequences flanking the recombination site are swapped; and non-crossovers, in which no swap occurs (Fig. 1). Previous studies[2] of several test regions had shown that crossovers and non-crossovers associated with gene conversion occur with similar frequencies. Although this suggested that crossovers and non-crossovers are alternative outcomes of a common pathway, subsequent studies indicated that they are produced by distinct molecular mechanisms[5].

Mancera and co-workers' data[1] reveal an unanticipated feature of genome-wide recombination patterns, namely the existence of regions (about 1% of the genome — an amount significantly greater than expected by chance) where the ratios of crossovers to non-crossovers differ substantially from the genome-wide average. The identification of hotspots where either non-crossovers or crossovers predominate, if confirmed, will provide regions where each type of recombination can be studied in relative isolation.

The existence of non-crossover hotspots also has implications for the design and interpretation of studies that use linkage disequilibrium (the disproportionate occurrence of certain gene combinations) to determine genetic association. If non-crossover hotspots exist, they would create 'holes' of low linkage disequilibrium within linkage-disequilibrium blocks, and copies of genes near such hotspots would be difficult to track relative to outside markers.

The authors' findings also reveal new features of the spacing between recombination events. It has been well documented (and confirmed by the present study) that crossovers show positive interference — that is, crossover at a locus reduces the likelihood of a second crossover nearby. This results in a more uniform inter-crossover spacing than is expected by chance. Previous studies that analysed thousands of tetrads found no evidence for positive interference between a crossover and a non-crossover[6,7]. By contrast, Mancera and co-workers' analysis of the spacing between non-crossovers and crossovers reveals a modest yet significant level of positive interference. This suggests that the spacing between recombination events is controlled, in part, at a very early step in recombination, perhaps at the time of formation of the double-strand break itself.

These findings, although intriguing, must be interpreted with caution. In particular, it is likely that a high marker density, which is required for high-resolution scoring of recombination, has the collateral consequence of altering the outcome of some events. Sequence differences between parental chromosomes at densities similar to those in this study have been shown to substantially alter the outcome of meiotic recombination[8], most likely by causing the disassembly of intermediates of inter-homologue recombination and redirecting events towards multiple exchange or genetically invisible recombination between sister chromatids. So, although Mancera and colleagues' data are probably representative of some 'real world' meioses — where hybrids with similarly high densities of sequence polymorphism are occasionally found — the above concerns should be kept in mind when considering the implications for recombination mechanisms.

Cautions and cavils aside, this fascinating study opens new doors in the study of meiotic recombination. For years to come, these data will be fertile material for analysis, and the insights gained from this approach in budding yeast should prompt others to develop similar methods to analyse recombination processes genome-wide in other organisms. ■

Michael Lichten is in the Laboratory of Biochemistry and Molecular Biology, Center for Cancer Research, National Cancer Institute, Bethesda, Maryland 20892-4260, USA.
e-mail: lichten@helix.nih.gov

1. Mancera, E., Bourgon, R., Brozzi, B., Huber, W. & Steinmetz, L. M. *Nature* **454,** 479–485 (2008).
2. Fogel, S., Mortimer, R., Lusnak, K. & Tavares, F. *Cold Spring Harb. Symp. Quant. Biol.* **43,** 1325–1341 (1979).
3. Wei, W. *et al. Proc. Natl Acad. Sci. USA* **104,** 12825–12830 (2007).
4. Pan, J. & Keeney, S. *PLoS Biol.* **5,** e333 (2007).
5. Bishop, D. K. & Zickler, D. *Cell* **117,** 9–15 (2004).
6. Mortimer, R. K. & Fogel, S. in *Mechanisms in Recombination* (ed. Grell, R. F.) 263–275 (Plenum, 1974).
7. Malkova, A. *et al. Genetics* **168,** 49–63 (2004).
8. Borts, R. H. & Haber, J. E. *Science* **237,** 1459–1465 (1987).

ECOLOGY

Forest air conditioning

F. I. Woodward

During the growing season, with photosynthesis at its peak, leaf temperatures remain constant over a wide latitudinal range. This is a finding that overturns a common assumption and has various ramifications.

The wood laid down as tree rings is rich in environmental information. The rings provide measures of annual growth. And the oxygen isotopic composition (^{18}O and ^{16}O) of wood cellulose provides estimates of historical temperature and humidity[1] — take a piece of wood, extract the cellulose, measure the ratio of ^{18}O to ^{16}O, slot the observations into a model of factors that favour one isotope over the other, and out pops a measure of ambient temperature and humidity.

But things are not so simple, say Helliker and Richter (page 511 of this issue)[2]. They have found that the oxygen-isotope ratios in cellulose collected across 50° of northern latitude, ranging from subtropical to boreal forest ecosystems, indicate that growing-season leaf temperatures are virtually constant. This latitudinal range has a 15 °C difference in growing-season temperatures. The more-or-less constant leaf temperatures (21.4 ± 2.2 °C) not only belie the assumption that leaf temperatures are the same as ambient temperatures, but also mean that humidity reconstructions will yield much lower values for cooler climates than would otherwise be expected and higher values for warmer climates.

The ratio of the ^{18}O and ^{16}O isotopes in wood cellulose is determined by the isotopic exchanges in leaves when they are synthesizing the sucrose from which cellulose is produced[3] (Fig. 1). The ratio is largely determined by the periods of maximum photosynthesis (around midday and around midsummer of the growing season), when the greatest amounts of sucrose are made before its incorporation into cellulose. The temperature-sensitive isotopic composition of precipitation and the rate of leaf evaporation (transpiration) both control the ^{18}O and ^{16}O ratio[1,3]. Determining ambient temperature and humidity directly from the ^{18}O:^{16}O composition of wood cellulose[1] depends on the simplifying assumption that leaves are at ambient temperature, and that the relative humidity of the air directly controls evaporation. This convenient assumption is highly unlikely to apply at times of maximum photosynthesis. The leaf temperature will

long-standing issues in meiotic recombination. I will focus on three: where recombination occurs; the type of recombination that occurs; and how recombination events are spaced along individual chromosomes.

Meiotic recombination is initiated by self-inflicted breaks in the double-stranded DNA; these breaks are then repaired by recombination with a homologous chromosome (Fig. 1). Genome-wide maps of meiotic double-strand breaks in budding yeast have indicated that most breaks occur in 'hotspots' that are distributed throughout the genome, with an average inter-hotspot distance of less than 10,000 bases[4]. Mancera and colleagues' recombination map[1] closely parallels these maps of double-strand breaks, with recombination occurring at almost all regions. In fact, the only truly recombination-cold regions of the yeast genome are associated with repeated sequences. These include mobile genetic elements, the DNA that encodes ribosomal RNA and sequences close to some chromosome ends.

The authors also investigated the relationship between the two outcomes of inter-homologue recombination: crossovers, where parental sequences flanking the recombination site are swapped; and non-crossovers, in which no swap occurs (Fig. 1). Previous studies[2] of several test regions had shown that crossovers and non-crossovers associated with gene conversion occur with similar frequencies. Although this suggested that crossovers and non-crossovers are alternative outcomes of a common pathway, subsequent studies indicated that they are produced by distinct molecular mechanisms[5].

Mancera and co-workers' data[1] reveal an unanticipated feature of genome-wide recombination patterns, namely the existence of regions (about 1% of the genome — an amount significantly greater than expected by chance) where the ratios of crossovers to non-crossovers differ substantially from the genome-wide average. The identification of hotspots where either non-crossovers or crossovers predominate, if confirmed, will provide regions where each type of recombination can be studied in relative isolation.

The existence of non-crossover hotspots also has implications for the design and interpretation of studies that use linkage disequilibrium (the disproportionate occurrence of certain gene combinations) to determine genetic association. If non-crossover hotspots exist, they would create 'holes' of low linkage disequilibrium within linkage-disequilibrium blocks, and copies of genes near such hotspots would be difficult to track relative to outside markers.

The authors' findings also reveal new features of the spacing between recombination events. It has been well documented (and confirmed by the present study) that crossovers show positive interference — that is, crossover at a locus reduces the likelihood of a second crossover nearby. This results in a more uniform inter-crossover spacing than is expected by chance. Previous studies that analysed thousands of tetrads found no evidence for positive interference between a crossover and a non-crossover[6,7]. By contrast, Mancera and co-workers' analysis of the spacing between non-crossovers and crossovers reveals a modest yet significant level of positive interference. This suggests that the spacing between recombination events is controlled, in part, at a very early step in recombination, perhaps at the time of formation of the double-strand break itself.

These findings, although intriguing, must be interpreted with caution. In particular, it is likely that a high marker density, which is required for high-resolution scoring of recombination, has the collateral consequence of altering the outcome of some events. Sequence differences between parental chromosomes at densities similar to those in this study have been shown to substantially alter the outcome of meiotic recombination[8], most likely by causing the disassembly of intermediates of inter-homologue recombination and redirecting events towards multiple exchange or genetically invisible recombination between sister chromatids. So, although Mancera and colleagues' data are probably representative of some 'real world' meioses — where hybrids with similarly high densities of sequence polymorphism are occasionally found — the above concerns should be kept in mind when considering the implications for recombination mechanisms.

Cautions and cavils aside, this fascinating study opens new doors in the study of meiotic recombination. For years to come, these data will be fertile material for analysis, and the insights gained from this approach in budding yeast should prompt others to develop similar methods to analyse recombination processes genome-wide in other organisms. ■

Michael Lichten is in the Laboratory of Biochemistry and Molecular Biology, Center for Cancer Research, National Cancer Institute, Bethesda, Maryland 20892-4260, USA.
e-mail: lichten@helix.nih.gov

1. Mancera, E., Bourgon, R., Brozzi, B., Huber, W. & Steinmetz, L. M. *Nature* **454,** 479–485 (2008).
2. Fogel, S., Mortimer, R., Lusnak, K. & Tavares, F. *Cold Spring Harb. Symp. Quant. Biol.* **43,** 1325–1341 (1979).
3. Wei, W. *et al. Proc. Natl Acad. Sci. USA* **104,** 12825–12830 (2007).
4. Pan, J. & Keeney, S. *PLoS Biol.* **5,** e333 (2007).
5. Bishop, D. K. & Zickler, D. *Cell* **117,** 9–15 (2004).
6. Mortimer, R. K. & Fogel, S. in *Mechanisms in Recombination* (ed. Grell, R. F.) 263–275 (Plenum, 1974).
7. Malkova, A. *et al. Genetics* **168,** 49–63 (2004).
8. Borts, R. H. & Haber, J. E. *Science* **237,** 1459–1465 (1987).

ECOLOGY

Forest air conditioning

F. I. Woodward

During the growing season, with photosynthesis at its peak, leaf temperatures remain constant over a wide latitudinal range. This is a finding that overturns a common assumption and has various ramifications.

The wood laid down as tree rings is rich in environmental information. The rings provide measures of annual growth. And the oxygen isotopic composition (^{18}O and ^{16}O) of wood cellulose provides estimates of historical temperature and humidity[1] — take a piece of wood, extract the cellulose, measure the ratio of ^{18}O to ^{16}O, slot the observations into a model of factors that favour one isotope over the other, and out pops a measure of ambient temperature and humidity.

But things are not so simple, say Helliker and Richter (page 511 of this issue)[2]. They have found that the oxygen-isotope ratios in cellulose collected across 50° of northern latitude, ranging from subtropical to boreal forest ecosystems, indicate that growing-season leaf temperatures are virtually constant. This latitudinal range has a 15 °C difference in growing-season temperatures. The more-or-less constant leaf temperatures (21.4 ± 2.2 °C) not only belie the assumption that leaf temperatures are the same as ambient temperatures, but also mean that humidity reconstructions will yield much lower values for cooler climates than would otherwise be expected and higher values for warmer climates.

The ratio of the ^{18}O and ^{16}O isotopes in wood cellulose is determined by the isotopic exchanges in leaves when they are synthesizing the sucrose from which cellulose is produced[3] (Fig. 1). The ratio is largely determined by the periods of maximum photosynthesis (around midday and around midsummer of the growing season), when the greatest amounts of sucrose are made before its incorporation into cellulose. The temperature-sensitive isotopic composition of precipitation and the rate of leaf evaporation (transpiration) both control the ^{18}O and ^{16}O ratio[1,3]. Determining ambient temperature and humidity directly from the ^{18}O:^{16}O composition of wood cellulose[1] depends on the simplifying assumption that leaves are at ambient temperature, and that the relative humidity of the air directly controls evaporation. This convenient assumption is highly unlikely to apply at times of maximum photosynthesis. The leaf temperature will

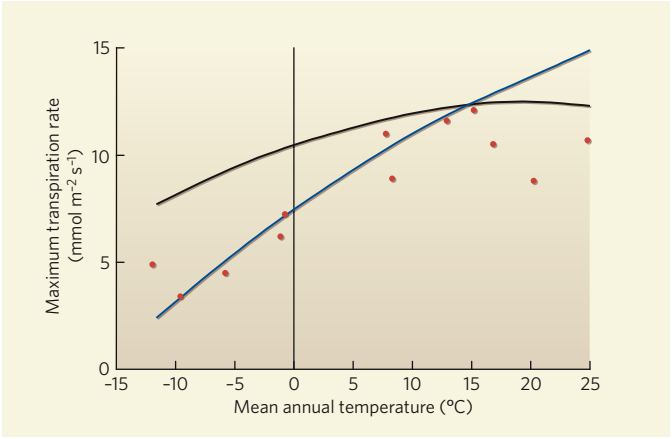
Box 1 | A test of constant leaf temperature

Helliker and Richter[2] provide climatic, isotopic and latitudinal data that can be used to test their conclusions by considering the energy balance of leaves in a canopy.

Convection is one way that leaves exchange energy with the ambient air; another is by evaporation of leaf water (transpiration). The rate of transpiration depends on the difference between leaf and air temperatures, and on the difference in humidity between the surrounding air and the leaf interior (assumed to be saturated at leaf temperature).

When leaf temperature equals air temperature, all energy exchange is assumed to occur by transpiration (black line in the figure). When the leaf temperature is constant and differs from the air temperature, energy exchange occurs by both convection and transpiration (blue line), and so transpiration rates are generally less than when leaf and air temperatures are equal. Observations of maximum rates of canopy transpiration[4,6] (red dots) can be compared with these expectations; all the data points are from forested sites, except the two at the lowest temperatures.

Constancy of leaf temperature[2] fits well with these measurements up to mean annual temperatures of 15 °C. Beyond that point, where the two curves cross, however, leaf temperatures of 21.4 °C will be less than the growing-season air temperatures. In consequence, leaves will be gaining heat from the surrounding air, further increasing transpiration, a response not observed at these forest sites[4]. **F.I.W.**



then exceed air temperature, and the gradient for evaporation is the difference in humidity between saturated air in the leaf and ambient air, which is quite different from the relative humidity of the air.

The approach adopted by Helliker and Richter[2] was to reject the assumption that leaves are at ambient temperature, and to calculate leaf temperatures using site-specific climatic data, the isotopic composition of precipitation and the ^{18}O:^{16}O composition of wood cellulose. The surprising outcome is that the calculated leaf temperatures are virtually constant across all of the sites sampled, contrasting with studies[1] that show large changes in calculated temperature with latitude. So, for example, during the growing season, leaf temperatures in a boreal forest will be on average 10 °C warmer than air temperatures, and the gradient for evaporation may be as much as twice that calculated under the assumption that air and leaf temperatures are equal.

This amount of temperature elevation is quite unexpected, especially for coniferous trees of the boreal zone — the leaves of these trees are needle-like and readily exchange energy with the surrounding air by convection, so that when isolated they closely follow air temperature. The explanation for the effect calculated by Helliker and Richter, as they themselves propose, is that in a forest the leaves are not isolated; rather, they are frequently bunched together in tight canopies (as anyone who has tried to stroll through mature boreal forest, or even a Sitka spruce plantation, will know). These dense and compact structures minimize the rate of convective energy exchange. Leaf temperatures are raised above ambient predominantly as a result of this canopy effect, rather than through a leaf-driven process.

Testing Helliker and Richter's conclusions is difficult, as leaf temperatures of the most actively photosynthesizing leaves in a forest canopy are not measured on a regular basis. An alternative test, which I have carried out, can be achieved by calculating the energy balance of leaves in a canopy over a wide range of growing-season temperatures, as described in Box 1. From this, one can conclude that the constancy in leaf temperature fits well with observations up to the point where the calculated leaf temperature equals the growing-season temperature. The data provided by Helliker and Richter imply that, beyond this point, leaf temperatures are lower than growing-season temperatures. As a consequence, leaves would be gaining heat from the surrounding air, a response not observed at selected forest sites[4].

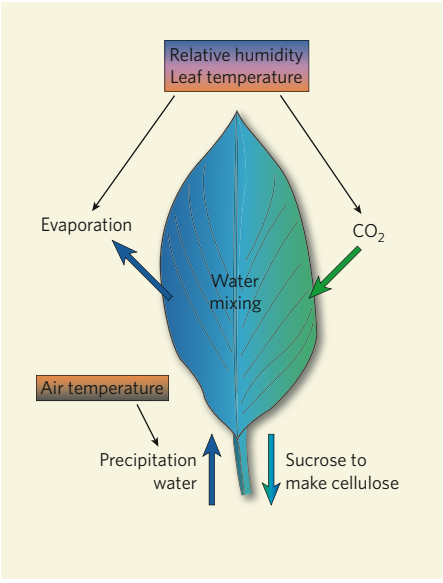


Figure 1 | Factors determining the ^{18}O:^{16}O ratio in wood cellulose. The ratio depends on differential discrimination against ^{18}O and ^{16}O in the leaf, and on the isotopic composition of water supplied by precipitation[2,3]. Evaporation (transpiration), which depends on the relative humidity of the air and on leaf temperature, favours the lighter isotope, and so the leaf becomes enriched in ^{18}O. The ^{18}O:^{16}O ratio of water in the sucrose produced by photosynthesis, which is a substrate for cellulose production, depends on a mix of enriched ^{18}O at the sites of evaporation and unenriched precipitation water. This ratio is influenced by the rate of evaporation and where in the leaf this mixing occurs. ^{18}O and ^{16}O from CO_2 exchange mixes completely with water before sucrose is synthesized.

A result of Helliker and Richter's study is that climatic reconstructions using oxygen isotopes in tree-ring cellulose will require reanalysis that takes into account the effects of leaf temperature and relative humidity on the ^{18}O:^{16}O ratio. Global vegetation models that assume that mean air temperature is the temperature for calculating both transpiration and photosynthesis will likewise need to be reviewed. Observations[4] show that the air temperature for maximum rates of carbon dioxide uptake increases by only 8 °C over a 35 °C change in mean annual temperature, indicating that the temperature response of photosynthesis is conservative, as expected from the constancy of leaf temperature.

Leaf structure and physiology show rather modest variations globally[5]. One cause may be the relative uniformity of leaf temperature — at least in fully forested vegetation. The fact that vegetation canopy rather than leaf morphology dominates temperature control in the forests sampled by Helliker and Richter suggests the need for greater emphasis on understanding how the canopy responds to climate change, and to global warming in particular. ■

F. I. Woodward is in the Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK.
e-mail: f.i.woodward@sheffield.ac.uk

1. Burk, R. L. & Stuiver, M. *Science* **211,** 1417–1419 (1981).
2. Helliker, B. R. & Richter, S. L. *Nature* **454,** 511–514 (2008).
3. Barbour, M. M. & Farquhar, G. D. *Plant Cell Environ.* **23,** 473–485 (2000).
4. AmeriFlux http://public.ornl.gov/ameriflux/data-access-select.shtml
5. Wright, I. J. *et al. Nature* **428,** 821–827 (2004).
6. Schaeffer, S. M., Williams, D. G. & Goodrich, D. C. *Agric. For. Meteorol.* **105,** 257–270 (2000).

HORIZONS

Life, logic and information

Paul Nurse

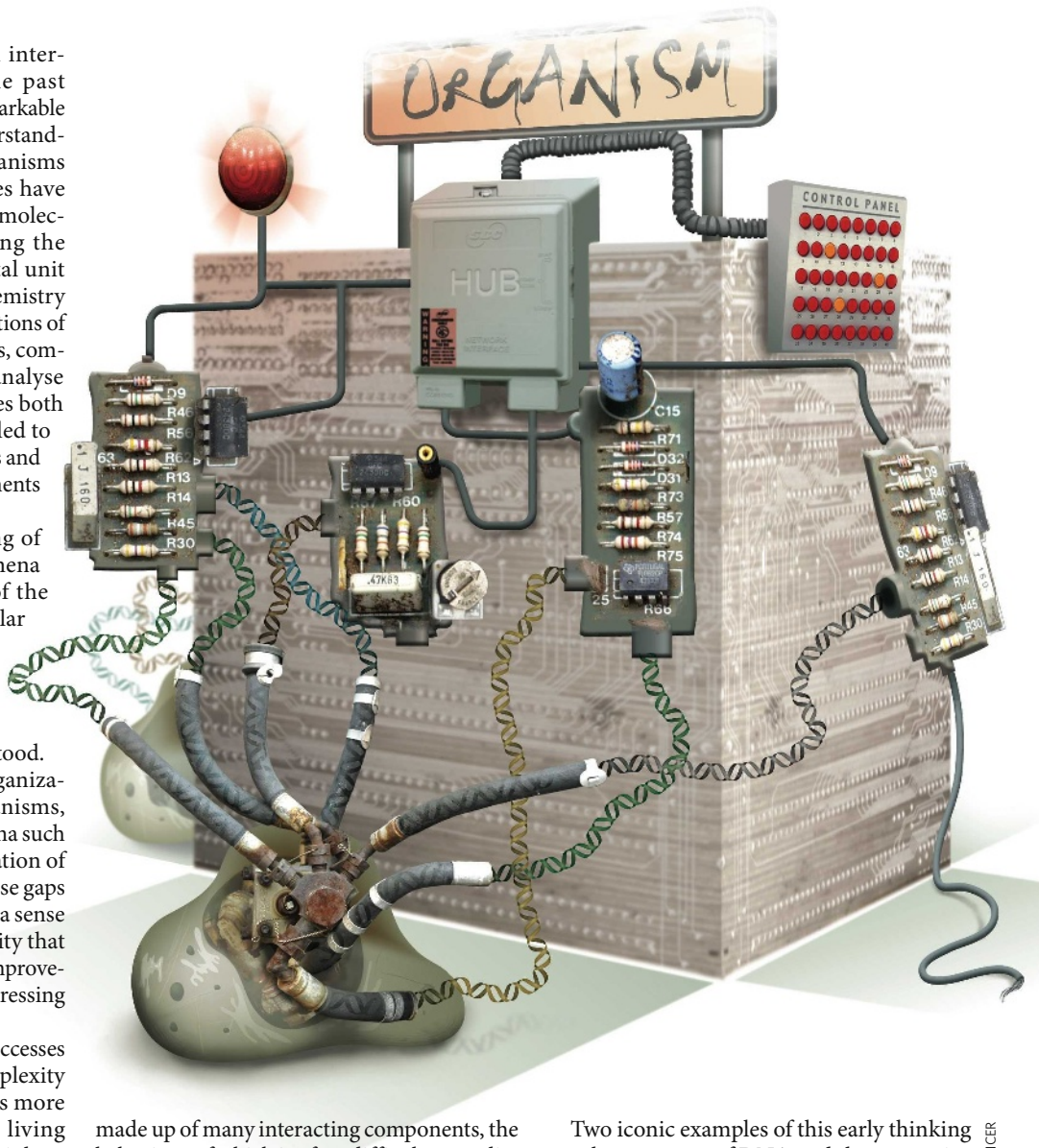
Focusing on information flow will help us to understand better how cells and organisms work.

Biology stands at an interesting juncture. The past decades have seen remarkable advances in our understanding of how living organisms work. These advances have been built mostly on molecular biology: applying the ideas that the gene is the fundamental unit of biological information and that chemistry provides effective mechanistic explanations of biological processes. These approaches, combined with an increasing ability to analyse highly complex biomolecular mixtures both qualitatively and quantitatively, have led to our present good understanding of cells and organisms and to significant improvements in our knowledge of human disease.

But comprehensive understanding of many higher-level biological phenomena remains elusive. Even at the level of the cell, phenomena such as general cellular homeostasis and the maintenance of cell integrity, the generation of spatial and temporal order, inter- and intracellular signalling, cell 'memory' and reproduction are not fully understood.

This is also true for the levels of organization seen in tissues, organs and organisms, which feature more complex phenomena such as embryonic development and operation of the immune and nervous systems. These gaps in our knowledge are accompanied by a sense of unease in the biomedical community that understanding of human disease and improvements in disease management are progressing too slowly.

One reason for this is that our past successes have led us to underestimate the complexity of living organisms. We need to focus more on how information is managed in living systems and how this brings about higher-level biological phenomena. There should be a concerted programme to investigate this, which will require both the development of the appropriate languages to describe information processing in biological systems and the generation of more effective methods to translate biochemical descriptions into the functioning of the logic circuits that underpin biological phenomena.



N. SPENCER

Living organisms are complex systems made up of many interacting components, the behaviour of which is often difficult to predict and so is prone to unexpected outcomes. Systems analyses of living organisms have used a variety of biochemical and genetic interaction traps with the emphasis on identifying the components and describing how these interact with each other. These approaches are essential but need to be supplemented by more investigation into how living systems gather, process, store and use information, as was emphasized at the birth of molecular biology.

Two iconic examples of this early thinking are the structure of DNA and the transcriptional regulation of the *lac* operon. The DNA double helix is beautiful not only because it is an elegant structure but because that structure reveals that DNA can act as a digital information storage device that can be precisely copied. Similarly, the mechanism of the *lac* operon (a set of nucleotides that regulates the metabolism of lactose) can be described in terms of molecular interactions between DNA, protein and metabolites. But these interactions make

sense only when they are translated into a negative feedback loop that processes information about the level of lactose in the environment to regulate the rate of *lac* operon transcription.

This type of thinking needs to be embraced more comprehensively in all studies of living processes. We need to describe the molecular interactions and biochemical transformations that take place in living organisms, and then translate these descriptions into the logic circuits that reveal how information is managed. This analysis should not be confined to the flow of information from gene to protein, but should also be applied to all functions operating in cells and organisms, including chemical interactions and transformations as well as physical phenomena, such as electrical signalling and mechanical processes.

Information management

The study of cells is likely to be particularly effective for this programme because the cell is the simplest entity that shows complex biological phenomena. Furthermore, model cellular systems, such as bacteria and yeasts, developing eggs of worms and flies, frog-egg extracts and mammalian cells, provide a range of powerful complementary genetic, genomic and biochemical experimental approaches.

Given the conservation of many processes, the model eukaryotic systems have the added advantage of being relevant to human cells. The aim should be to analyse cells more effectively with the intention of then applying those approaches to more difficult organismal problems and to human disease. Two phases of work are required for such a programme: to describe and catalogue the logic circuits that manage information in cells, and to simplify analysis of cellular biochemistry so that it can be linked to the logic circuits.

For the first phase, the logic circuits that operate within cells need to be broken down into the individual segments that carry out specific computational functions. I shall call these segments 'logic modules'. One example of such a module is the negative feedback loop, which often operates in a homeostatic manner. Another example is the positive feedback loop, which can generate irreversible switch behaviour from one state to another. Combinations of modules will produce more sophisticated outcomes: for example, reversible toggle switches, timers and oscillators.

The behaviour of the outputs from modules will be influenced by the shapes of the response curves embedded within them, with the outputs generated depending on whether, for example, the curves are linear, hyperbolic or sigmoidal. Modules could act as a short-term memory device, as seen in a G protein locked in a GTP-bound state, or as a long-term digital memory device as in the case of DNA. The identification of the logic modules used in cellular systems will allow a catalogue to be generated that defines the logic 'tool-kit' that is available to cells.

A useful analogy is an electronic circuit. Representations of such circuits use symbols to define the nature and function of the electronic components used. They also describe the logic relationships between the components, making it clear how information flows through the circuit. A similar conceptualization is required of the logic modules that make up the circuits that manage information in cells.

The initial identification of the logic modules operating in cells requires detailed biochemical descriptions of the interactions between different molecular components. Knowledge of the rate constants and strengths of interactions allows models to be built and differential equations to be generated and solved. If constraints exist as to what sorts of modules and linkages can generate effective and robust behaviours, then fewer possibilities will need to be considered. The tool-kit of modules and of the linkages between them that operate in cells may thus be limited, reducing the complexity of the problem that has to be solved.

"Studies at higher system levels are likely to inform those at the simpler level of the cell and vice versa."

Knowledge of which modules are operational and how these are linked into circuits will help us to understand the flow of information. We need to know how information is gathered from various sources, from the environment, from other cells and from the short- and long-term memories in the cell; how that information is integrated and processed; and how it is then either used, rejected or stored for later use. The aim is to describe how information flows through the modules and brings about higher-level cellular phenomena, investigations that may well require the development of new methods and languages to describe the processes involved.

The next phase will be to simplify the analysis of the cellular biochemistry and link it with the logic modules. Key to this is determining which molecules interact with each other. This analysis is well under way with the application of various interaction-trapping approaches, such as two-hybrid methods, protein purification followed by mass spectrometry, and genetic screens for synthetic lethality. A further approach will be the systematic cataloguing of the position of fluorescently tagged proteins in living cells to identify which proteins are near to each other and how that proximity may change over time. These spatial and temporal descriptions of molecules within living cells should simplify the analysis by defining a limited set of cellular spatial and temporal 'domains' that need to be considered. All these data will then need to be organized into databases, relating different cell types and model systems.

The next step is difficult, as it involves the mapping of molecular interactions and biochemical functions onto the logic modules, in effect linking the cellular chemistry tool-kit with the logic tool-kit. The success of this mapping will depend on whether there are sufficient regularities between specific logic modules and specific interacting molecules, at least at some level of probability.

Such regularities may not exist if natural selection has recruited many different components from the chemical tool-kit to generate specific examples of the logic tool-kit. However, there may be sufficient regularities to make this mapping possible. The fact that life on Earth generally uses nucleic acids as digital information-storage devices, gives some cause for optimism. Another example may be protein kinases and phosphatases that act antagonistically, which behave like switches.

As we learn more about how molecules interact to generate logic modules it may become less necessary to know the details of the rate constants and the molecular concentrations and to solve the differential equations that they generate. If detailed modelling reveals that certain molecules wired together in particular ways are often associated with specific modules, then it might become possible to predict some behaviours without having precise measurements of the variables involved. Simply knowing which molecular components are present and how they are linked together might be sufficient to speculate about which logic module is in operation. If this is the case, then the module can be considered as a black box and it may be necessary to concentrate only on *in vivo* measurements of key inputs into and outputs from the black box to confirm that the logic module is behaving in the expected manner.

Analysis in practice

How could such a programme work in practice? First the higher-level cellular phenomenon of interest has to be identified. Examples of such processes include chemotaxis, mating, signalling and aspects of cellular reproduction. One approach would then be to mutationally saturate the phenomenon by use of forward genetics and genome-wide deletion collections to identify as many of the genes involved as possible. Application of standard bioinformatic procedures would link the genes identified with specific biochemical and molecular functions. Identifying which molecules interact with each other, and how, can be established by use of the interaction trap, and by spatial and temporal cellular domain databases.

So far this approach is relatively conventional. The next steps will be to use the databases described above to determine the probability that specific components of the chemical tool-kit are associated with a particular logic module. Finally, the modules will be linked together into a complete circuit, allowing outputs to be predicted so that the functioning of the circuit can be translated into a narrative of information flow to describe

how the cellular phenomenon works.

What issues might we expect to encounter if this programme is adopted? One important consideration is that because the logic modules and circuits are combined into networks, an understanding of how such networks operate in cells will be crucial. Complex networks have been well analysed in other spheres of human activity. For example, transportation networks such as flight routes and connections are often found to have diverse numbers of linkages between hubs in the network such that some hubs become crucial because they are highly connected to many other hubs. Network analysts call such networks 'scale-free'. It seems that biological networks derived from genetic, protein–protein and transcriptional interaction studies are also often scale-free. So far, analysis has suggested that these hubs are likely to be ancient in origin and so arose early in evolution.

It is important to realize that unlike simpler networks such as those seen in transportation systems, linkages between hubs in cellular networks will not all be of a similar physical and logic type. Some will represent stable physical interactions and others will reflect more transient biochemical reactions. Furthermore, the logic consequences will vary, either negative or positive in action, for example. In the future it will be necessary to use representations that capture more effectively the different linkages connecting hubs in biological systems. Biological networks are also more flexible and fluid, and can reconnect and reassemble in different ways to generate alternative networks with changed outcomes. The language used to properly represent biological networks will need to accommodate these variations in logic structures.

Dynamic signals

Another interesting feature of logic circuits in biological systems is the roles that temporal organization or dynamics may have. Signalling pathways within or between cells have generally been thought of as linear sequences that lead to on/off switches. An analogy for such a sequence is a railway signal that results in only one of two outcomes, a stop or a go signal. If dynamics is introduced into signalling pathways, richer behaviours can emerge. For example, if signals are pulsed down a pathway and the changing outputs are monitored, much more complex information can be transmitted.

A metaphor here would be the use of the Morse code and the telegraph to communicate messages. Pulses of information sent along the telegraph generate a code for letters and as a consequence sentences can be communicated. This converts the same signalling pathway from a simple on/off switch to a device that can transfer, for example, the works of Shakespeare. It is likely that dynamics has been exploited more generally in the evolution of biological systems for signalling purposes, allowing the communication of more complex information.

Spatial organization of signalling pathways within cells will also enrich behaviours, with different outcomes being possible in different regions of the cell depending on the spatial context of the input and output signals. Logic circuits can also give rise to behaviours that generate spatial organization, as in the case of Alan Turing's reaction-diffusion equations. Because cells are extended in space, the spatial organization generated by logic circuits will contribute to spatial order within the cell, for example by acting as position-locating mechanisms during the generation of cellular form.

Finally, we need to take account of the biological origins of the logic circuits and networks that operate in cells. Because natural selection operates on pre-existing living organisms, novelties will initially arise as add-ons to systems already in existence, almost guaranteeing some redundancy. Thus, man-made machines, which are generally intelligently designed, will differ from the logic machines found in life. Living machines are not intelligently designed and will often be redundant and overly complex.

"Living machines are not intelligently designed and will often be redundant and overly complex."

We should anticipate these differences and be prepared for the additional complexity to be found in the logic circuits that manage information in cells. Lessons will also be learned from the higher levels of biological organization seen in communities of individuals, in ecological systems and during evolutionary change. The principles and rules that underpin how information is managed may share similarities at these different levels even though their elements are completely different. Studies at higher system levels are thus likely to inform those at the simpler level of the cell and vice versa.

I have suggested that cells and experimentally amenable model systems should form the major part of this programme at this point in time, but ultimately what we learn with these simpler biological systems needs to be applied to more complex multicellular organisms and to humans if we are to fully understand organismal biology and improve treatment of human disease. Part of the problem of shifting these approaches to organisms will be one of scale, of having to deal with more genes, more involved structures and more complex phenomena. It will also be necessary to take full account of ecological and environmental interactions as well as the evolutionary context of the organism under study. In addition, we will have to develop methodologies to properly investigate intact living organisms, including humans in both the healthy and the diseased state. Particularly important for this work will be the development of high-resolution sensitive imaging procedures to monitor biomolecules in real time and in space. This is the return to whole-organism and human physiology that many have argued is long overdue, but with a renewed emphasis on the logic of life and the management of information.

Programme requirements

What is required that is not already generally in place to pursue this programme effectively? Perhaps the most pressing need is to develop the appropriate theoretical approaches to analyse the management of information flow and to investigate the logic systems that are responsible for that flow.

I see this best being developed not as a 'big science' project but by individual scientists working alone and together in small interactive workshop groups meeting on a regular basis. The groups will need to be multidisciplinary, including information theorists, mathematicians, physicists, chemists and computer scientists working closely with experimental biologists who have good biological intuition and who can communicate with members of the other disciplines. Different workshop groups could interact with each other through digital conversations to share ideas.

The training of advanced undergraduate and graduate biologists also needs to shift in its emphasis. The separation of molecular and cell biologists from those that study organism biology, ecology and evolution has weakened biomedical research, and the emphasis on learning large numbers of facts in molecular- and cell-biology courses and during medical training has reduced the necessary exposure to the ideas central to biology.

Time needs to be made during education to expose biomedical scientists to other scientific disciplines to ensure good communication between biologists and other disciplines so that theory is always well embedded in biological facts and experiments. Placing a greater emphasis on ideas during teaching and training will have the added advantage of attracting excellent students to the whole biological and biomedical research endeavour.

Success in the programme will require sophisticated databases that can manage different types of data from a range of experimental systems that can be used to generate connections and handle probabilities of outcomes. New experimental techniques are required to allow better *in vivo* analysis of living systems with sophisticated imaging for real-time experiments. The analyses will also need to develop beyond single-organism studies in closely defined unchanging laboratory conditions, and move towards more complex ecological circumstances working with societies of organisms in changing environments. ■

Paul Nurse is at the Rockefeller University, 1230 York Avenue, New York, New York 10065, USA.
e-mail: nurse@mail.rockefeller.edu

Acknowledgements I would like to thank members of my laboratory and Emily Nurse for their helpful comments on this article.

Cover illustration
A polarized light micrograph of crystals of uric acid, which causes the inflammatory condition gout. (Courtesy of R. J. Green/SPL)

Editor, *Nature*
Philip Campbell

Publishing
Samia Mantoura
Claudia Banks

Insights Editor
Ritu Dhand

Production Editor
Davina Dadley-Moore

Senior Art Editor
Martin Harrison

Art Editor
Nik Spencer

Sponsorship
Amélie Pequignot

Production
Jocelyn Hilton

Marketing
Elena Woodstock

Editorial Assistant
Alison McGill

INFLAMMATION

Inflammation is the body's immediate response to damage to its tissues and cells by pathogens, noxious stimuli such as chemicals, or physical injury. Acute inflammation is a short-term response that usually results in healing: leukocytes infiltrate the damaged region, removing the stimulus and repairing the tissue. Chronic inflammation, by contrast, is a prolonged, dysregulated and maladaptive response that involves active inflammation, tissue destruction and attempts at tissue repair. Such persistent inflammation is associated with many chronic human conditions and diseases, including allergy, atherosclerosis, cancer, arthritis and autoimmune diseases.

The processes by which acute inflammation is initiated and develops are well defined, but much less is known about the causes of chronic inflammation and the associated molecular and cellular pathways. This Insight highlights recent advances in our knowledge of the exogenous and endogenous inducers of chronic inflammation, as well as the inflammatory mediators and cells that are involved. We hope that these articles will contribute to a better understanding of inflammatory responses, and ultimately result in the design of more effective therapies for the numerous debilitating diseases with a chronic inflammatory component.

Ursula Weiss, Senior Editor

REVIEWS

HYPOTHESIS

REVIEW

nature insight

NATURE|Vol 454|24 July 2008|doi:10.1038/nature07201

Origin and physiological roles of inflammation

Ruslan Medzhitov[1]

Inflammation underlies a wide variety of physiological and pathological processes. Although the pathological aspects of many types of inflammation are well appreciated, their physiological functions are mostly unknown. The classic instigators of inflammation — infection and tissue injury — are at one end of a large range of adverse conditions that induce inflammation, and they trigger the recruitment of leukocytes and plasma proteins to the affected tissue site. Tissue stress or malfunction similarly induces an adaptive response, which is referred to here as para-inflammation. This response relies mainly on tissue-resident macrophages and is intermediate between the basal homeostatic state and a classic inflammatory response. Para-inflammation is probably responsible for the chronic inflammatory conditions that are associated with modern human diseases.

Inflammation is an adaptive response that is triggered by noxious stimuli and conditions, such as infection and tissue injury[1,2]. Considerable progress has been made in understanding the cellular and molecular events that are involved in the acute inflammatory response to infection and, to a lesser extent, to tissue injury. In addition, the events that lead to localized chronic inflammation, particularly in chronic infections and autoimmune diseases, are partly understood. Much less is known, however, about the causes and mechanisms of systemic chronic inflammation, which occurs in a wide variety of diseases, including type 2 diabetes and cardiovascular diseases. These chronic inflammatory states do not seem to be caused by the classic instigators of inflammation: infection and injury. Instead, they seem to be associated with the malfunction of tissue: that is, with the homeostatic imbalance of one of several physiological systems that are not directly functionally related to host defence or tissue repair (Fig. 1).

It is generally thought that a controlled inflammatory response is beneficial (for example, in providing protection against infection), but it can become detrimental if dysregulated (for example, causing septic shock). Thus, the pathological inflammatory state is assumed to have a physiological counterpart. However, whereas the physiological rationale of infection-induced inflammation is clear, many other types of inflammatory response are only known in pathological settings, and there is no clear understanding of their physiological counterparts. It is not even clear whether there is any physiological counterpart for some inflammatory conditions, such as gout and obesity. Whether or not there are physiological counterparts for all inflammatory conditions, an important piece of the puzzle is missing from the current understanding of the inflammatory process. The standard view of inflammation as a reaction to infection or injury might need to be expanded to account for the inflammatory processes induced by other types of adverse conditions.

Regardless of the cause, inflammation presumably evolved as an adaptive response for restoring homeostasis. Therefore, the origin of inflammatory responses is perhaps best understood in this broader context. Here I discuss some of the physiological roots of inflammation as an adaptive response to tissue malfunction or homeostatic imbalance. Different types of inflammatory inducer, including altered cellular and tissue states, are discussed in this context.

Overview of the inflammatory response

At a basic level, the acute inflammatory response triggered by infection or tissue injury involves the coordinated delivery of blood components (plasma and leukocytes) to the site of infection or injury[1,2]. This response has been characterized best for microbial infections (particularly bacterial infections), in which it is triggered by receptors of the innate immune system, such as Toll-like receptors (TLRs) and NOD (nucleotide-binding oligomerization-domain protein)-like receptors (NLRs)[3]. This initial recognition of infection is mediated by tissue-resident macrophages and mast cells, leading to the production of a variety of inflammatory mediators, including chemokines, cytokines, vasoactive amines, eicosanoids and products of proteolytic cascades. The main and most immediate effect of these mediators is to elicit an inflammatory exudate locally: plasma proteins and leukocytes (mainly neutrophils) that are normally restricted to the blood vessels now gain access, through the postcapillary venules, to the extravascular tissues at the site of infection (or injury). The activated endothelium of the blood vessels allows selective extravasation of neutrophils while preventing the exit of erythrocytes. This selectivity is afforded by the inducible ligation of endothelial-cell selectins with integrins and chemokine receptors on leukocytes, which occurs at the endothelial surface, as well as in the extravascular spaces (where newly deposited plasma proteins form a provisional matrix for the binding of leukocyte integrins)[4]. When they reach the afflicted tissue site, neutrophils become activated, either by direct contact with pathogens or through the actions of cytokines secreted by tissue-resident cells. The neutrophils attempt to kill the invading agents by releasing the toxic contents of their granules, which include reactive oxygen species (ROS) and reactive nitrogen species, proteinase 3, cathepsin G and elastase[5]. These highly potent effectors do not discriminate between microbial and host targets, so collateral damage to host tissues is unavoidable[6].

A successful acute inflammatory response results in the elimination of the infectious agents followed by a resolution and repair phase, which

[1]Howard Hughes Medical Institute and Department of Immunobiology, Yale University School of Medicine, TAC S-669, 300 Cedar Street, New Haven, Connecticut 06510, USA.

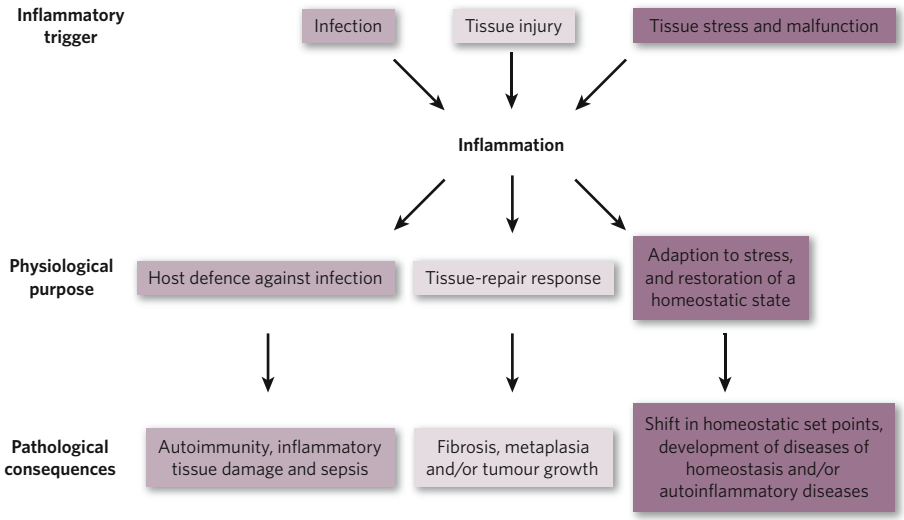


Figure 1 | Causes, and physiological and pathological outcomes, of inflammation. Depending on the trigger, the inflammatory response has a different physiological purpose and pathological consequences. Of the three possible initiating stimuli, only infection-induced inflammation is coupled with the induction of an immune response.

is mediated mainly by tissue-resident and recruited macrophages[7]. The switch in lipid mediators from pro-inflammatory prostaglandins to lipoxins, which are anti-inflammatory, is crucial for the transition from inflammation to resolution. Lipoxins inhibit the recruitment of neutrophils and, instead, promote the recruitment of monocytes, which remove dead cells and initiate tissue remodelling[7]. Resolvins and protectins, which constitute another class of lipid mediator, as well as transforming growth factor-β and growth factors produced by macrophages, also have a crucial role in the resolution of inflammation, including the initiation of tissue repair[7,8].

If the acute inflammatory response fails to eliminate the pathogen, the inflammatory process persists and acquires new characteristics. The neutrophil infiltrate is replaced with macrophages, and in the case of infection also with T cells. If the combined effect of these cells is still insufficient, a chronic inflammatory state ensues, involving the formation of granulomas and tertiary lymphoid tissues[2,9]. The characteristics of this inflammatory state can differ depending on the effector class of the T cells that are present. In addition to persistent pathogens, chronic inflammation can result from other causes of tissue damage such as autoimmune responses (owing to the persistence of self antigens) or undegradable foreign bodies. Unsuccessful attempts by macrophages to engulf and destroy pathogens or foreign bodies can lead to the formation of granulomas, in which the intruders are walled off by layers of macrophages, in a final attempt to protect the host[1,2].

It should be noted that the mechanisms of infection-induced inflammation are understood far better than are those of other inflammatory processes. It is unclear how applicable knowledge of infection-induced inflammation is to other types of inflammation. Indeed, although infection-induced inflammation is vital, it might be a special case. The mechanisms of systemic chronic inflammatory states in general are poorly understood, but it is clear that they do not seem to fit the classic pattern of transition from acute inflammation to chronic inflammation.

The inflammatory 'pathway'

The inflammatory response is coordinated by a large range of mediators that form complex regulatory networks. To dissect these complex networks, it is helpful to place these signals into functional categories and to distinguish between inducers and mediators of inflammation. Inducers are the signals that initiate the inflammatory response. They activate specialized sensors, which then elicit the production of specific sets of mediators. The mediators, in turn, alter the functional states of tissues and organs (which are the effectors of inflammation) in a way that allows them to adapt to the conditions indicated by the particular inducer of inflammation. Thus, a generic inflammatory 'pathway' consists of inducers, sensors, mediators and effectors, with each component determining the type of inflammatory response (Fig. 2a and Table 1). In the following sections, these four pathway components are discussed in turn.

Inducers and sensors of inflammation

Inducers of inflammation can be exogenous or endogenous (Fig. 2b).

Exogenous inducers of inflammation

Exogenous inducers can be classified into two groups: microbial and non-microbial. There are, in turn, two classes of microbial inducer: pathogen-associated molecular patterns (PAMPs) and virulence factors. The first class of microbial inducer, PAMPs, is a limited and defined set of conserved molecular patterns that is carried by all microorganisms of a given class (whether pathogenic or commensal)[10]. PAMPs are defined in the sense that the host has evolved a corresponding set of receptors (known as pattern-recognition receptors) that detect their presence.

The second class of microbial inducer comprises a variety of virulence factors and is therefore restricted to pathogens. In contrast to PAMPs, they are not sensed directly by dedicated receptors. Instead, the effects of their activity, particularly their adverse effects on host tissues, are responsible for triggering the inflammatory response. Typical activities of virulence factors can be detected by specialized sensors. For example, the pore-forming exotoxins produced by Gram-positive bacteria are detected by the NALP3 (NACHT-, leucine-rich-repeat- and pyrin-domain-containing protein) inflammasome, which is sensitive to the efflux of K^+ ions that results from pore formation[11]. Similarly, the proteolytic activity of proteases produced by helminths is sensed by basophils by an unknown sensor[12]. Notably, this sensing mechanism can be inadvertently activated by functional mimics, so allergens that are proteases can trigger the pathway that is usually induced by helminths[12]. An alternative way of sensing virulence activity is non-specific and even more indirect, through detecting the effects on cell death and tissue damage. In this case, the actual inducers of the inflammatory response are endogenous products of damaged cells and tissues. Importantly, the inflammatory responses that are induced by these two sensing mechanisms of virulence activity differ in their specificity, because the former is characteristic of pathogens (and in some cases, pathogen classes), but the latter is not. These inflammatory responses are likely to have different characteristics, and it will be interesting to investigate whether they result in distinct physiological and pathological outcomes.

It should be emphasized that microbial inducers of inflammation are not necessarily derived from pathogens. Commensal bacteria provide an important source of inflammation inducers that are detected by TLRs[13]. The activation of TLRs by these bacteria is actively suppressed by multiple mechanisms. An example of this is the lethal TLR-dependent inflammation that develops in mice that lack A20, one of the crucial negative regulators of TLR signalling[14].

Exogenous inducers of inflammation that are of non-microbial origin include allergens, irritants, foreign bodies and toxic compounds[1]. Certain allergens are detected because they mimic the virulence activity of parasites (as mentioned earlier); others can act as irritants on the mucosal epithelia. The inflammatory response induced by both types of allergen is largely similar because defence against parasites and environmental irritants relies on expulsion and clearance mediated by the mucosal epithelia. The sensors for allergens are largely unknown.

Foreign bodies are indigestible particles that either are too large to be phagocytosed or cause phagosomal membrane damage in macrophages. Silica and asbestos particles are notorious examples of foreign bodies that elicit an inflammatory response. Their large size and resistance to removal, as well as a lack of self markers (such as CD47) that are normally present on autologous cells and prevent their phagocytosis (by engaging inhibitory receptors), point to an abnormal occurrence in the tissues. The 'missing self' recognition presumably triggers a 'phagocytic reflex' in macrophages, but the large size or the shape of foreign particles results in 'frustrated phagocytosis': that is, a phagocytic cup is formed but cannot close to form a phagosome. If a foreign body is too large for a phagocytic cup to be formed, the macrophage forms a granuloma around this body instead. The sensor that triggers this reaction in macrophages is unknown. In some cases, macrophages can fuse with each other to form 'giant cells' that encapsulate the foreign body. The encapsulation of foreign objects is an ancient defensive strategy, which is also found in *Drosophila melanogaster*, in which lamellocytes (macrophage-like cells) encapsulate parasitoid wasp eggs to protect the host[15]. Regardless of whether a foreign body is too large to be phagocytosed or disrupts the phagosomal membrane, when a macrophage encounters foreign bodies, the NALP3 inflammasome (a sensor) is activated[16].

Endogenous inducers of inflammation

Endogenous inducers of inflammation are signals produced by stressed, damaged or otherwise malfunctioning tissues (discussed later). The identity and characteristics of these signals are poorly defined. But they probably belong to various functional classes according to the nature and the degree of tissue anomalies on which they report.

One common (but not universal) theme in detecting acute tissue injury is the sensing of the desequestration of cells or molecules that are normally kept separate in intact cells and tissues. The sequestration of these components (for example, ligands and their receptors, or enzymes and their activators or substrates) is afforded by the various types of compartmentalization that occur in normal tissues. Important examples are sequestration bounded by cellular membranes (especially the plasma membrane), basement membranes, the surface epithelium and the vascular endothelium.

During necrotic cell death, for example, the integrity of the plasma membrane is disrupted, resulting in the release of certain cellular constituents, including ATP, K^+ ions, uric acid, HMGB1 (high-mobility group box 1 protein) and several members of the S100 calcium-binding protein family (S100A8, S100A9 and S100A12)[17,18]. ATP binds to purinoceptors (including $P2X_7$) at the surface of macrophages, resulting in K^+ ion efflux, and can cooperate with other signals to activate the NALP3 inflammasome[11]. ATP also activates nociceptors (which are sensory receptors), thereby reporting tissue injury to the nervous system[19]. HMGB1 and S100A12 engage the receptor RAGE (advanced glycation end-product-specific receptor; also known as AGER), which (at least in the case of HMGB1) cooperates with TLRs to induce an inflammatory response[20,21]. S100A8 and S100A9 signal through TLR4 (ref. 22). It should be noted that, although intracellular proteins are thought to be passively released when the plasma membrane of necrotic cells is disrupted, numerous intracellular proteins can be secreted by way of a non-canonical (endoplasmic-reticulum–Golgi-independent) pathway. A recent study has shown that this non-canonical secretion is mediated by activated caspase 1, implying that the secretion is regulated by inflammasomes[23]. In light of this finding, it will be necessary to examine whether inflammatory intracellular proteins are passively released from necrotic cells or secreted by way of this caspase-1-dependent mechanism. These two possibilities are mutually exclusive for a given cell, because necrotic cells are metabolically inactive, whereas caspase-1-dependent secretion is an ATP-driven process. If caspase 1 is responsible for the secretion of intracellular proteins with inflammatory activities, this will shed a different light on the role of intracellular inflammatory proteins in initiating inflammation, as well as on the role of necrotic cell death. The prime example in this case is HMGB1, which has been shown to be secreted by macrophages stimulated with the TLR4 ligand lipopolysaccharide[24], apparently in the absence of necrotic cell death, suggesting that the non-canonical caspase-1-dependent secretory pathway might be involved.

In intact tissues, epithelial cells and mesenchymal cells are normally separated from each other by the basement membrane, and the disruption of this barrier results in 'unscheduled' epithelial–mesenchymal interactions. These interactions indicate the presence of tissue damage and consequently initiate tissue-repair responses, but how these abnormal interactions are sensed is poorly understood. The surface epithelia separate the internal compartments from the external environment. In organs, such as the intestine, that are colonized by commensal microorganisms, the disruption of the epithelial barrier gives commensal microorganisms access to the TLRs on macrophages that reside in the lamina propria, resulting in TLR-mediated induction of tissue-repair responses in the intestine[13,25]. In sterile organs with an epithelial lining, the desequestration of some non-microbial luminal components might have a similar role. Another remarkable example of the use of a desequestration strategy is the separation of the growth factor heregulin (also known as neuregulin 1) from its receptors (ERBB2, ERBB3 and ERRB4) in the airway epithelium[26]. The tight junctions of the intact polarized epithelium separates heregulin, which is apically expressed, from its receptors, which are basolaterally expressed, thereby preventing their interaction. On epithelial injury, heregulin gains access to its receptors and initiates a tissue-repair response[26].

Finally, damage to the vascular endothelium allows plasma proteins and platelets to gain access to extravascular spaces[4]. A key plasma-derived regulator of inflammation, the Hageman factor (also known as factor XII), becomes activated by contact with collagen and other components of the extracellular matrix (ECM). Activated Hageman factor acts as a sensor of vascular damage and initiates the four proteolytic cascades that generate inflammatory mediators: the kallikrein–kinin cascade, the coagulation cascade, the fibrinolytic cascade and the complement cascade[1]. Platelets are also activated by contact with collagen and produce various inflammatory mediators, including thromboxanes and serotonin[1].

The endogenous inducers that have been discussed so far are involved in acute inflammatory responses to tissue injury. Another class of endogenous inducer is more relevant to chronic inflammatory conditions. This class of inducer includes crystals of monosodium urate and calcium pyrophosphate dihydrate, AGEs (advanced glycation end products) and oxidized lipoproteins (such as high-density lipoproteins and low-density lipoproteins). The formation of such crystals is facilitated in certain connective tissues, which provide an appropriate surface for crystal nucleation[17]. The formation of monosodium urate and calcium pyrophosphate dihydrate crystals in the joints and periarticular tissues, for example, is responsible for the inflammatory conditions known as gout and pseudogout, respectively[17]. When these crystals reach a certain size, they are detected by macrophages and treated in essentially the same way as foreign bodies (discussed earlier). Phagocytosis of these particles triggers the activation of the NALP3 inflammasome and subsequently the production of caspase-1 substrates, including members of the interleukin 1 (IL-1) family[16,27].

AGEs are products of the non-enzymatic glycation of long-lived proteins, such as collagen[28]. These products can result in the crosslinking of the proteins they are attached to, leading to gradual deterioration of the function of these proteins. In addition, AGEs are recognized by their receptor, RAGE, which has inflammatory activity either alone[21] or in combination with TLRs[29]. AGEs can accumulate under hyperglycaemic and pro-oxidative conditions, including type 1 and type 2 diabetes, and ageing[28]. ROS, produced by phagocytes, also have a role in converting

high-density lipoproteins and low-density lipoproteins into inflammatory signals by oxidizing their lipid and protein components[30].

Another group of endogenous inducers of inflammation consists of breakdown products of the ECM that are generated during tissue malfunction or damage. The best-studied component of the ECM in this context is the glycosaminoglycan hyaluronate. In normal conditions, hyaluronate is present as an inert high-molecular-weight polymer. Tissue injury promotes its breakdown into low-molecular-weight fragments, which are inflammatory, activating TLR4 and promoting a tissue-repair response[31]. This conversion is also thought to be ROS dependent[32]. Thus, several endogenous pathways that initiate the inflammatory response depend on ROS.

The list of endogenous inducers of inflammation is growing, but the scientific literature on this subject contains many discrepancies. This is largely due to the technical difficulties that are associated with characterizing this class of signal. A common reason for incorrectly identifying a factor as an inducer results from contamination of recombinant proteins with traces of microbial ligands for TLRs or NOD proteins. More importantly, many endogenous inducers of inflammation presumably exert the appropriate activity *in vivo* only when present in certain combinations and perhaps only in the context of malfunctioning or damaged tissues. For example, ischaemia (local lack of blood supply), hypoxia, increased concentrations of ROS and altered ECM components are all commonly associated with tissue damage or malfunction but are not reproduced in tissue-culture conditions, which are commonly characterized by supra-physiological nutrient and oxygen concentrations.

In addition to the inducers associated with infection and tissue damage, there is probably another, currently unidentified, class of inducer that triggers the inflammatory response in tissues that are malfunctioning or are under stress. These signals report on the homeostatic status of tissues and induce adaptive changes that involve some hallmarks of the classic inflammatory response (discussed later).

Mediators and effectors of inflammation

Inducers of inflammation trigger the production of numerous inflammatory mediators, which in turn alter the functionality of many tissues and organs — the downstream effectors of the inflammatory pathway. Many of these inflammatory mediators have effects in common on the vasculature and on the recruitment of leukocytes. These mediators can be derived from plasma proteins or secreted by cells[1,2]. The cellular mediators can be produced by specialized leukocytes (particularly tissue-resident macrophages and mast cells) or by cells present in local tissues. Some mediators (such as histamine and serotonin) are preformed and stored in the granules of mast cells, basophils and platelets. Others are preformed and circulate as inactive precursors in the plasma. The plasma concentration of these mediators can increase markedly as a result of increased secretion of the precursors by hepatocytes during the acute-phase response. Other mediators are produced directly in response to appropriate stimulation by inducers of inflammation.

Inflammatory mediators can be classified into seven groups according to their biochemical properties[1,2]: vasoactive amines, vasoactive peptides, fragments of complement components, lipid mediators, cytokines, chemokines and proteolytic enzymes.

First, vasoactive amines (histamine and serotonin) are produced in an all-or-none manner when mast cells and platelets degranulate. They have complex effects on the vasculature, causing increased vascular permeability and vasodilation, or vasoconstriction, depending on the context. The immediate consequences of their release by mast cells can be highly detrimental in sensitized organisms, resulting in vascular and respiratory collapse during anaphylactic shock.

Second, vasoactive peptides can be stored in an active form in secretory vesicles (for example, substance P) or generated by proteolytic processing of inactive precursors in the extracellular fluid (for example, kinins, fibrinopeptide A, fibrinopeptide B and fibrin degradation products). Substance P is released by sensory neurons and can itself cause mast-cell degranulation. Other vasoactive peptides are generated through proteolysis by the Hageman factor, thrombin or plasmin and cause vasodilation and increased vascular permeability (either directly or by inducing the release of histamine from mast cells). As mentioned earlier, the Hageman factor has a key role in coordinating these responses, and it functions as both a sensor of vascular damage and an inducer of inflammation. The Hageman factor activates the kallikrein–kinin cascade, and the main product of this cascade, bradykinin, affects the vasculature, as well as having a potent pro-algesic (pain-stimulating) effect. Pain sensation has an important physiological role in inflammation by alerting the organism to the abnormal state of the damaged tissue.

Third, the complement fragments C3a, C4a and C5a (also known as anaphylatoxins) are produced by several pathways of complement activation. C5a (and to a lesser extent C3a and C4a) promote granulocyte and monocyte recruitment and induce mast-cell degranulation, thereby affecting the vasculature.

Fourth, lipid mediators (eicosanoids and platelet-activating factors) are derived from phospholipids, such as phosphatidylcholine, that are present in the inner leaflet of cellular membranes. After activation by intracellular Ca^{2+} ions, cytosolic phospholipase A_2 generates arachidonic acid and lysophosphatidic acid, the precursors of the two classes of lipid mediator listed above, from phosphatidylcholine. Arachidonic acid is metabolized to form eicosanoids either by cyclooxygenases (COX1 and COX2), which generate prostaglandins and thromboxanes, or by lipoxygenases, which generate leukotrienes and lipoxins[2]. The prostaglandins PGE_2 and PGI_2, in turn, cause vasodilation, and PGE_2 is also hyperalgesic and a potent inducer

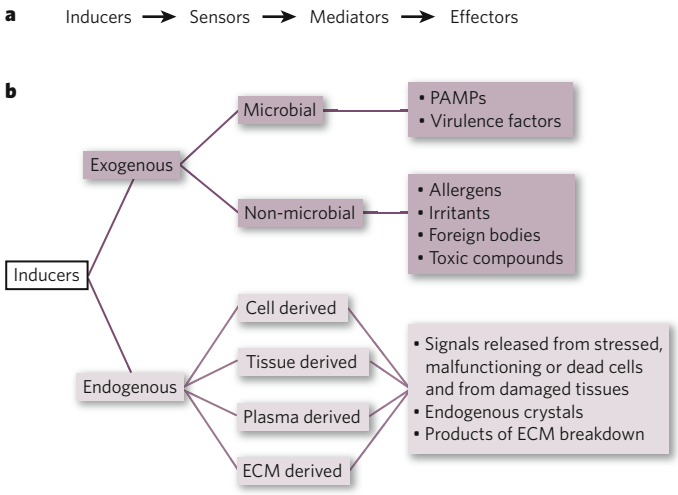


Figure 2 | The inflammatory pathway. a, A generic inflammatory pathway consists of inducers, sensors, mediators and effectors. Table 1 provides examples of each component for several inflammatory pathways. **b**, Inducers of inflammation can be classified as exogenous or endogenous, and these two groups can be further classified as shown. ECM, extracellular matrix; PAMP, pathogen-associated molecular pattern.

Table 1 | Examples of inflammatory pathways

Inducer	Sensor	Mediator	Effectors
Lipopolysaccharide	TLR4	TNF-α, IL-6 and PGE_2	Endothelial cells, hepatocytes, leukocytes, the hypothalamus, and others
Allergens	IgE	Vasoactive amines	Endothelial cells and smooth muscle cells
Monosodium urate crystals and calcium pyrophosphate dihydrate crystals	NALP3	IL-1β	Endothelial cells, hepatocytes, leukocytes, the hypothalamus, and others
Collagen	Hageman factor	Bradykinin	Endothelial cells and smooth muscle cells

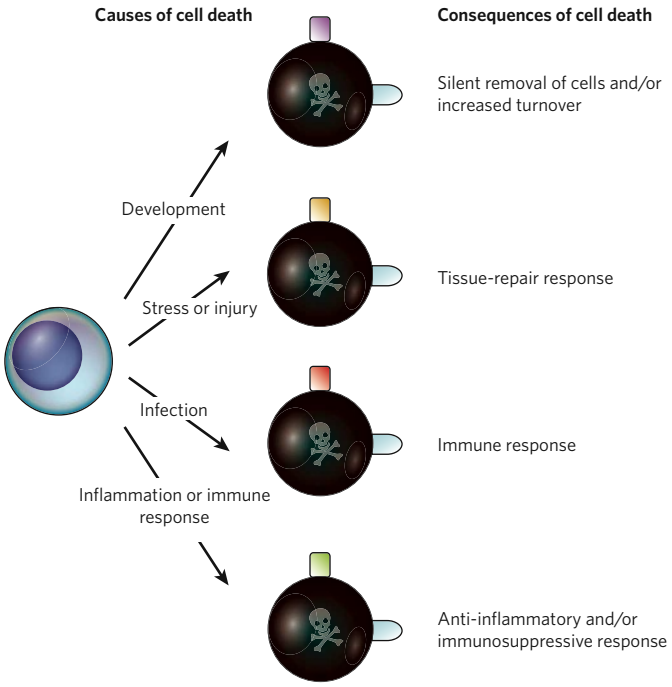


Figure 3 | Cell death and its consequences. Apoptotic cells display the lipid phosphatidylserine (blue) at the plasma membrane, resulting in their recognition and, subsequently, phagocytosis by macrophages. In addition, apoptotic cells probably produce other signals (coloured rectangles) that determine the outcome of their recognition by macrophages, but the type of signal is likely to depend on the cause of cell death.

of fever[33]. Lipoxins (and dietary ω3-fatty-acid-derived resolvins and protectins) inhibit inflammation and promote resolution of inflammation, and tissue repair[8]. The second class of lipid mediator, platelet-activating factors, are generated by the acetylation of lysophosphatidic acid and activate several processes that occur during the inflammatory response, including recruitment of leukocytes, vasodilation and vasoconstriction, increased vascular permeability and platelet activation[1,2].

Fifth, inflammatory cytokines (tumour-necrosis factor-α (TNF-α), IL-1, IL-6 and many others) are produced by many cell types, most importantly by macrophages and mast cells. They have several roles in the inflammatory response, including activation of the endothelium and leukocytes and induction of the acute-phase response.

Sixth, chemokines are produced by many cell types in response to inducers of inflammation. They control leukocyte extravasation and chemotaxis towards the affected tissues.

Seventh, several proteolytic enzymes (including elastin, cathepsins and matrix metalloproteinases) have diverse roles in inflammation, in part through degrading ECM and basement-membrane proteins. These proteases have important roles in many processes, including host defence, tissue remodelling and leukocyte migration.

It should be noted that it is unclear to what extent the nature of an inflammatory trigger dictates the type of mediator induced. In addition, many (but not all) mediators not only have direct effects on target tissues but also themselves induce the production of additional mediators. It will be important to understand the logic underlying this hierarchy of mediators.

The effectors of an inflammatory response are the tissues and cells, the functional states of which are specifically affected by the inflammatory mediators. Responsiveness to certain inflammatory mediators (such as TNF-α and IL-1) is almost ubiquitous, although these mediators have distinct effects in different tissue and cell types. Although the most obvious effect of inflammatory mediators is to induce the formation of an exudate (through their effects on the vasculature and on leukocyte migration), many inflammatory mediators have other, equally important, effects on neuroendocrine and metabolic functions and on the maintenance of tissue homeostasis in general[34]. These functions of inflammatory mediators reflect a more general role for inflammation in the control of tissue homeostasis and in adaptation to noxious conditions.

Homeostatic control through stress response and adaptation

Homeostatic control mechanisms ensure that internal environmental parameters (such as glucose and oxygen concentrations) are maintained within an acceptable range near a certain set point[35]. Abnormal conditions can cause a deviation in some parameters beyond the normal homeostatic range, resulting in either an acute stress response that affords a transient adaptation to the new conditions or a more sustained adaptive change that involves a shift in the relevant set points. In a general sense, acute inflammation and chronic inflammation are different types of adaptive response that are called into action when other homeostatic mechanisms are either insufficient or not competent.

The inflammatory response is commonly thought to operate during severe disturbances of homeostasis, such as infection, tissue injury and the presence of foreign bodies or irritants. However, infection and injury are at the extreme end of a spectrum of conditions that can trigger inflammation, and they trigger responses of the highest magnitude (which is why these are the best known and characterized inflammatory responses). More generally, an inflammatory response is presumably engaged whenever tissue malfunctions are detected. These types of inflammatory response are likely to be more common but of lower magnitude than the classic inflammatory responses induced by infection or injury. The nature and the degree of tissue malfunction will influence whether the inflammatory responses is detectable using common biomarkers. Such tissue alterations can range from mild tissue-specific malfunctions to massive injury. Accordingly, the magnitude of the inflammatory response can differ markedly. Very mild stress might be handled by tissue-resident cells (mainly macrophages and mast cells), whereas more extensive malfunctions or damage might require additional leukocytes to be recruited and plasma proteins to be delivered locally. These latter effects (in response to extensive malfunction or damage) are those of a classic inflammatory response. Unlike the signals that report infection and injury, the signals that report tissue stress and malfunction, and the molecular sensors that detect these signals, are largely unknown.

Whatever the cause of the inflammatory response, its 'purpose' is to remove or sequester the source of the disturbance, to allow the host to adapt to the abnormal conditions and, ultimately, to restore functionality and homeostasis to the tissue. If the abnormal conditions are transient, then a successful acute inflammatory response returns the system to the basal homeostatic set points. If, by contrast, the abnormal conditions are sustained, then an ongoing inflammatory state shifts the system to different set points, as occurs during chronic inflammation. An adaptive change often provides short-term benefits; however, in a chronic phase, it can become maladaptive, as exemplified by a sustained decline in insulin sensitivity of the skeletal muscle or by squamous metaplasia of the respiratory epithelium. More specifically, a transient decrease in insulin sensitivity during acute inflammation would allow the redistribution of glucose from one of its major consumers (for example, skeletal muscle) to leukocytes and other cell types that can have an increased energy demand during infection and tissue repair. However, sustained insulin resistance in skeletal muscle can lead to type 2 diabetes. Likewise, squamous metaplasia can have a short-term benefit by protecting the respiratory tract from damage by irritants, but it results in a decline in respiratory function in the long term. Indeed, inducible adaptive changes generally occur at the expense of many other physiological processes and therefore cannot be sustained without adverse side effects caused by the decline in the affected functions. For example, the acute-phase response and oedema both have adaptive values during bacterial infections but occur at the expense of the normal functioning of many tissues. The marked increase in plasma protein concentration that occurs during the acute-phase response alters the oncotic pressure, which has many potential adverse effects on the circulatory system, and oedema causes local hypoxia by increasing the

distance between parenchymal cells and capillaries. The potential for adverse effects is intrinsic to any adaptive changes, whether these changes occur at the cellular, tissue or organismal level.

Because any tissue malfunction is initiated by changes at the cellular level, I now discuss the cellular alterations that might be associated with the initiation of an inflammatory response.

Cell states

Any cell can be in one of four possible states: basal, stressed, apoptotic or necrotic. A cell is in a basal state when conditions are normal, and this state is maintained by the availability of nutrients, oxygen and growth factors, and by attachment to other cells and/or the ECM. A change in any of the vital internal environmental parameters (temperature, osmolarity, oxygen, and so on) induces a stress response, which is in essence a cellular adaptation to the abnormal condition. If the change in a parameter is greater than the stress response can handle, the cell undergoes apoptosis. If the change is greater still, the cell undergoes necrosis. Developmentally controlled apoptosis occurs for reasons unrelated to tissue maintenance and is not discussed here.

Each of the four cellular states is regulated by specialized signalling pathways. Recent evidence indicates that even necrosis, previously considered to be an accidental and unplanned form of cell death, is regulated by dedicated genetic programs[36]. Importantly, the pathways that induce or maintain each of the states are 'wired' so that they inhibit the transition to the next state, as shown by the following examples. The basal state can be maintained by a prototypical insulin-like growth factor 1 (IGF1)-dependent cell-survival pathway, which inhibits the generic stress response regulated by FOXO transcription factors[37]. The stress-inducible nuclear factor-κB pathway inhibits apoptosis by engaging multiple mechanisms[38]. The key effectors of apoptosis — caspase 3, caspase 6 and caspase 7 — cleave and inactivate PARP (which is involved in DNA repair), thereby blocking PARP-dependent necrosis[36] (which results from the depletion of cytosolic NAD, a substrate of PARP[36,39]). Thus, the more desirable cellular state (which in descending order is basal, stressed, apoptotic, then necrotic) inhibits the transition to the next, less desirable, state until the transition is unavoidable. This fundamental property of cell-fate decisions has two important implications. First, it ensures that the transition occurs in a switch-like manner rather than gradually. This is important because it enables the cells to exhaust their attempts to stay in a more preferable state before making the transition to the next, less preferable, state. Second, each cellular state can express distinct sets of signals that report the cellular state, and the switch-like transition prevents the generation of mixed messages.

The state of cells and tissues is probably monitored mainly by tissue-resident macrophages (and, in some tissues, also by mast cells). When tissues are in the basal state, tissue-resident macrophages maintain tissue homeostasis by a variety of tissue-specific mechanisms. Tissue-resident macrophages constitute 10–15% of most tissues, and their functions extend beyond host defence and the removal of apoptotic cells[40,41]. Examples are control of the turnover of epithelial cells, regulation of the metabolic activity of adipocytes and remodelling of bone (which is carried out by osteoclasts)[40,41]. When tissues are in conditions of stress, or when they malfunction for other reasons, they might send a different set of signals to tissue-resident macrophages than those sent by tissues in the basal state. The tissue-resident macrophages, in turn, produce increased amounts, or different sets, of growth factors and other signals that are relevant for the particular tissue. When the stress or malfunction is extreme, the help provided by local macrophages might be insufficient, and the tissues might 'call for' the recruitment of additional macrophages. Thus, malfunctioning adipocytes in obese animals secrete the chemokine CC-chemokine ligand 2 (CCL2), which recruits more macrophages to the adipose tissue[42]. Hypoxic tissues produce the chemokine CXC-chemokine ligand 12 (CXCL12), which can also recruit macrophages[43]. There are also many other cases in which macrophages are recruited in a tissue-specific or condition-specific manner[44]. Furthermore, tissue-derived signals can control the activation state and type of recruited macrophage[44,45]. The main purpose of these interactions is to help the tissues to adapt to the stressful conditions and to restore their functionality. However, when these interactions are sustained or excessive, they can become maladaptive, as is evident from macrophages that have been recruited to sites of inflammation contributing to insulin resistance in adipose tissue. Moreover, the accessory function of macrophages (assisting the adaptation of tissues to stress conditions) can be exploited by tumour cells, which can recruit macrophages and use them as a source of growth factors, angiogenic factors and chemokines. Several examples of such exploitation by tumour cells have been documented and shown to have a crucial role in tumour progression and metastasis[46].

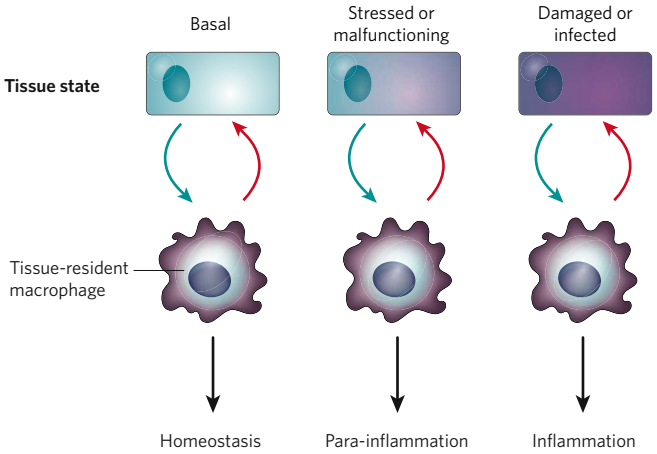


Figure 4 | Three modes of adaptation and maintenance of tissue homeostasis. The state of a tissue can range from basal, to stressed or malfunctioning, to damaged or infected, and each state is graded (as indicated by shading). The state affects the mode of maintenance of tissue homeostasis or adaptive response that is engaged by tissue-resident macrophages and, in some tissues, by other types of leukocyte. Blue arrows indicate signals that report the tissue state to macrophages; red arrows indicate macrophage-derived signals that control tissue adaptation. At one extreme of the range of responses is inflammation, which follows infection or tissue damage. By contrast, tissue stress or malfunction induces para-inflammation, which helps a tissue to adapt to the noxious conditions and restore tissue functionality. Dysregulated para-inflammation might be responsible for the chronic inflammatory states that are associated with many modern human diseases, such as type 2 diabetes and atherosclerosis.

If tissue malfunction or stress is excessive and adaptation is no longer possible, the cells die by apoptosis or necrosis. Infection and tissue injury are the most common contributors to this transition, but other insults also have this effect. These insults can be classified as 'inflammation inducers' (discussed earlier). In these cases, cell death is again monitored and interpreted by macrophages. In addition to the removal of apoptotic and necrotic cells, macrophages make one of several possible 'decisions', ranging from the silent removal of dead cells to the induction of an inflammatory response. Because necrotic cell death is generally associated with tissue damage, the outcome of necrotic-cell recognition by macrophages is usually an inflammatory response[17,36,47]. Apoptosis, by contrast, can occur for several reasons, and macrophages therefore need to be able to decipher the cause of death to take the appropriate actions. In this way, for each of the four (or more) situations in which apoptosis occurs, a different outcome is possible (Fig. 3).

First, during developmentally programmed apoptosis, dead cells are removed by macrophages without any additional consequences. Because this form of apoptosis is a normal part of development or cell turnover, no further action is needed (although in the case of cell turnover, macrophages might produce growth factors that promote cell proliferation, in order to replace dead cells).

Second, apoptosis induced by excessive stress or injury results in unscheduled cell loss, which needs to be compensated for by the generation of new cells of the same type. Therefore, on recognition of cells that have died prematurely, macrophages should (in general) induce a tissue-repair response.

Box 1 | Evolution of adaptive traits

Why is the inflammatory response closely associated with pathological conditions? To answer this question, it is worthwhile considering some basic principles of the evolution of adaptive traits[56].

It is often assumed that any given physiological process, as a product of evolution, has an adaptive value. However, some traits can evolve without being adaptive, owing to chance or constraints of an organism's history[57]. Similarly, evolution does not necessarily produce an optimal solution to every problem.

For the purpose of this discussion, it is useful to consider two situations. The first is that a particular characteristic (or a trait) can be adaptive (that is, beneficial) if it was selected for because it had a positive effect on the organism's fitness (and, ultimately, on the organism's reproductive success). It is important to note that such traits are adaptive in the conditions that were present when they evolved. If these conditions change sufficiently, the same traits can be maladaptive. Obesity and allergy are examples of maladaptive traits. The second situation is that a trait can be non-adaptive when it exists as a consequence of another (adaptive) trait but has no positive value of its own. Anaphylactic shock and tissue destruction by activated neutrophils are examples of non-adaptive traits. These non-adaptive traits were not selected for, because they are either neutral or detrimental with respect to an organism's fitness (in the conditions in which they evolved) but exist as an unavoidable consequence of an adaptive trait. The second situation is common and constitutes an evolutionary trade-off between the beneficial effects of adaptive traits and the detrimental effects of non-adaptive traits with which they are coupled. In the conditions in which the traits evolved, the beneficial effects must have outweighed the detrimental effects. But, again, a change in conditions can shift the balance in this trade-off, making the non-adaptive traits a substantial burden to the organism. This situation is particularly characteristic of young species, such as humans, that have not yet reached an evolutionary equilibrium. Many modern human diseases are the result of unbalanced trade-offs caused by marked changes in environmental conditions and lifestyles.

When considering the inflammatory process, there are many examples of adaptive and non-adaptive traits. However, for many chronic inflammatory processes, only the pathological (maladaptive) aspects are evident, and there is no understanding of their physiological (adaptive) counterparts, which are presumed to exist. Distinguishing between adaptive and non-adaptive characteristics is essential not only for gaining a deeper understanding of inflammation but also for developing efficient therapeutic strategies. For example, both adaptive and non-adaptive processes can contribute to disease symptoms (owing to inherent trade-offs), but interfering with the adaptive processes can, ultimately, make matters worse, whereas blocking the non-adaptive processes should be beneficial and should not cause adverse side effects.

Third, apoptosis induced by infection (including the caspase-1-dependent process, known as pyroptosis[48]) should switch the macrophages to a host-defence mode, thereby promoting the generation of an immune response.

Fourth, apoptosis induced by inflammatory or immune responses should have the opposite effect to infection-induced apoptosis. The recognition of apoptotic cells by macrophages, in this case, should result in the induction of anti-inflammatory and immunosuppressive pathways. This is exemplified by the recently described anti-inflammatory pathway controlled by the TAM family of receptor tyrosine kinases[49].

Macrophages recognize all apoptotic cells (regardless of how apoptosis is induced) by detecting phosphatidylserine at the surface of these cells[50,51]. The outcome of the various forms of apoptosis is probably determined, however, by additional signals, which are likely to be differentially produced by apoptotic cells that have died from different causes.

Para-inflammation

Cell states are discrete rather than graded, and transitions between these states occur in an all-or-none manner (as discussed earlier). By contrast, tissue states are graded: tissues can contain different numbers of dead cells, for example, or they can malfunction to different degrees. Accordingly, the adaptive response elicited by the tissues can take different forms depending on the degree of the problem that is experienced. Thus, in basal conditions, the tissues are maintained in a homeostatic state, in many cases with the help of the tissue-resident macrophages. In noxious conditions, tissues undergo stress and can malfunction. If the changes are considerable, then adaptation to the conditions requires the help of tissue-resident or recruited macrophages and might require small-scale delivery of additional leukocytes and plasma proteins, depending on the extent of the problem. This adaptive response has characteristics that are intermediate between basal and inflammatory states. It could be termed para-inflammation (*para-* being the Greek prefix for near).

Para-inflammatory responses are graded: at one extreme, they are close to the basal state, whereas, at the other, they start to transition into inflammation (Fig. 4). The induction of a para-inflammatory response does not require overt tissue injury or infection; instead, it is switched on by tissue malfunction, in order to restore tissue functionality and homeostasis. If tissue malfunction is present for a sustained period, para-inflammation can become chronic. Sustained malfunction can result from mutations or environmental factors. It can also be caused by the maladaptive traits that are responsible for modern human diseases (Box 1). Indeed, many chronic inflammatory diseases that are not caused by infection or injury seem to be associated with conditions that were not present during the early evolution of humans, including the continuous availability of high-calorie nutrients, a low level of physical activity, exposure to toxic compounds, and old age. The human diseases that are associated with these conditions — including obesity, type 2 diabetes, atherosclerosis, asthma and neurodegenerative diseases — are all characterized by chronic low-grade inflammation (para-inflammation), which might not have any physiological counterparts. Furthermore, the chronic para-inflammation that persists in these conditions can, in turn, contribute to further disease progression, in part because of changes in homeostatic set points (such as insulin sensitivity or blood pressure).

The inflammatory range proposed here is consistent with inflammation having a general physiological role in maintaining tissue homeostasis, monitoring tissue malfunction, and promoting adaptation to adverse conditions and dysfunctional states that the tissues cannot resolve by themselves. This idea is in accordance with the fact that many inflammatory mediators (including TNF-α, IL-6, CCL2 and prostaglandins) also have important homeostatic functions, for example in repair of tissues, control of metabolism, and regulation of the hypothalamus–pituitary axis[52–54]. Thus, inflammatory processes can extend the homeostatic capacity of the organism and complement the homeostatic controls provided by the endocrine system and the autonomic nervous system.

Conclusions

Inflammation is exceedingly complex and equally fascinating. It has a crucial role in mammalian physiology. Many components of the inflammatory response have been found in all vertebrates studied so far, and some forms of adaptive response to adverse conditions (including infection and injury) occur in all animals and plants. However, the complexity of vertebrates — which are characterized by having multiple renewable tissues — perhaps necessitated the development of a specialized adaptive and protective capacity that is provided by the inflammatory response.

This invention came at a price, however. The pathological potential of inflammation is unprecedented for a physiological process[55]. Although the destructive ability of infection-induced inflammation is understandably unavoidable, the pathogenic capacity of other types of inflammation is puzzling. A major unresolved problem is defining the normal physiological counterpart of the systemic chronic inflammatory state. The idea of para-inflammation might provide part of the answer. In a broader context, it will be necessary to understand better which pathological aspects of inflammation are the results of trade-offs with beneficial traits and which are maladaptive for other reasons. ■

1. Majno, G. & Joris, I. *Cells, Tissues and Disease* (Oxford Univ. Press, 2004).
2. Kumar, V., Cotran, R. S. & Robbins, S. L. *Robbins Basic Pathology* (Saunders, 2003).
3. Barton, G. M. A calculated response: control of inflammation by the innate immune system. *J. Clin. Invest.* **118,** 413–420 (2008).
4. Pober, J. S. & Sessa, W. C. Evolving functions of endothelial cells in inflammation. *Nature Rev. Immunol.* **7,** 803–815 (2007).
5. Nathan, C. Neutrophils and immunity: challenges and opportunities. *Nature Rev. Immunol.* **6,** 173–182 (2006).
6. Nathan, C. Points of control in inflammation. *Nature* **420,** 846–852 (2002).
7. Serhan, C. N. & Savill, J. Resolution of inflammation: the beginning programs the end. *Nature Immunol.* **6,** 1191–1197 (2005).
8. Serhan, C. N. Resolution phase of inflammation: novel endogenous anti-inflammatory and proresolving lipid mediators and pathways. *Annu. Rev. Immunol.* **25,** 101–137 (2007).
9. Drayton, D. L., Liao, S., Mounzer, R. H. & Ruddle, N. H. Lymphoid organ development: from ontogeny to neogenesis. *Nature Immunol.* **7,** 344–353 (2006).
10. Medzhitov, R. & Janeway, C. A. Jr Innate immunity: the virtues of a nonclonal system of recognition. *Cell* **91,** 295–298 (1997).
11. Mariathasan, S. *et al.* Cryopyrin activates the inflammasome in response to toxins and ATP. *Nature* **440,** 228–232 (2006).
This study shows that pore-forming exotoxins activate the NALP3 inflammasome.
12. Sokol, C. L., Barton, G. M., Farr, A. G. & Medzhitov, R. A mechanism for the initiation of allergen-induced T helper type 2 responses. *Nature Immunol.* **9,** 310–318 (2008).
13. Rakoff-Nahoum, S., Paglino, J., Eslami-Varzaneh, F., Edberg, S. & Medzhitov, R. Recognition of commensal microflora by Toll-like receptors is required for intestinal homeostasis. *Cell* **118,** 229–241 (2004).
14. Turer, E. E. *et al.* Homeostatic MyD88-dependent signals cause lethal inflammation in the absence of A20. *J. Exp. Med.* **205,** 451–464 (2008).
This paper shows that the activation of TLRs by commensal microorganisms can result in lethal inflammation in the absence of a negative regulator of TLR signalling.
15. Rizki, T. M. & Rizki, R. M. Lamellocyte differentiation in *Drosophila* larvae parasitized by *Leptopilina*. *Dev. Comp. Immunol.* **16,** 103–110 (1992).
16. Dostert, C. *et al.* Innate immune activation through Nalp3 inflammasome sensing of asbestos and silica. *Science* **320,** 674–677 (2008).
This study shows that environmental particles can trigger inflammation through the NALP3 inflammasome.
17. Rock, K. L. & Kono, H. The inflammatory response to cell death. *Annu. Rev. Pathol.* **3,** 99–126 (2008).
18. Bianchi, M. E. DAMPs, PAMPs and alarmins: all we need to know about danger. *J. Leukoc. Biol.* **81,** 1–5 (2007).
19. Julius, D. & Basbaum, A. I. Molecular mechanisms of nociception. *Nature* **413,** 203–210 (2001).
20. Park, J. S. *et al.* High mobility group box 1 protein interacts with multiple Toll-like receptors. *Am. J. Physiol. Cell Physiol.* **290,** C917–C924 (2006).
21. Hofmann, M. A. *et al.* RAGE mediates a novel proinflammatory axis: a central cell surface receptor for S100/calgranulin polypeptides. *Cell* **97,** 889–901 (1999).
22. Vogl, T. *et al.* Mrp8 and Mrp14 are endogenous activators of Toll-like receptor 4, promoting lethal, endotoxin-induced shock. *Nature Med.* **13,** 1042–1049 (2007).
23. Keller, M., Ruegg, A., Werner, S. & Beer, H. D. Active caspase-1 is a regulator of unconventional protein secretion. *Cell* **132,** 818–831 (2008).
This study shows that caspase 1 regulates the secretion of many cytosolic proteins by a non-canonical (ER–Golgi-independent) mechanism.
24. Chen, G. *et al.* Bacterial endotoxin stimulates macrophages to release HMGB1 partly through CD14- and TNF-dependent mechanisms. *J. Leukoc. Biol.* **76,** 994–1001 (2004).
25. Pull, S. L., Doherty, J. M., Mills, J. C., Gordon, J. I. & Stappenbeck, T. S. Activated macrophages are an adaptive element of the colonic epithelial progenitor niche necessary for regenerative responses to injury. *Proc. Natl Acad. Sci. USA* **102,** 99–104 (2005).
26. Vermeer, P. D. *et al.* Segregation of receptor and ligand regulates activation of epithelial growth factor receptor. *Nature* **422,** 322–326 (2003).
27. Martinon, F., Petrilli, V., Mayor, A., Tardivel, A. & Tschopp, J. Gout-associated uric acid crystals activate the NALP3 inflammasome. *Nature* **440,** 237–241 (2006).
This study shows that uric acid (urate) crystals induce inflammation by activating the NALP3 inflammasome.
28. Brownlee, M., Cerami, A. & Vlassara, H. Advanced glycosylation end products in tissue and the biochemical basis of diabetic complications. *N. Engl. J. Med.* **318,** 1315–1321 (1988).
This paper describes AGEs and their role in inflammation.
29. Yan, S. F. *et al.* The biology of RAGE and its ligands: uncovering mechanisms at the heart of diabetes and its complications. *Curr. Diab. Rep.* **7,** 146–153 (2007).
30. Navab, M. *et al.* Mechanisms of disease: proatherogenic HDL — an evolving field. *Nature Clin. Pract. Endocrinol. Metab.* **2,** 504–511 (2006).
31. Jiang, D. *et al.* Regulation of lung injury and repair by Toll-like receptors and hyaluronan. *Nature Med.* **11,** 1173–1179 (2005).
This study shows the protective role of TLR-induced tissue repair in sterile tissue injury.
32. Jiang, D., Liang, J. & Noble, P. W. Hyaluronan in tissue injury and repair. *Annu. Rev. Cell Dev. Biol.* **23,** 435–461 (2007).
33. Higgs, G. A., Moncada, S. & Vane, J. R. Eicosanoids in inflammation. *Ann. Clin. Res.* **16,** 287–299 (1984).
34. Turnbull, A. V. & Rivier, C. L. Regulation of the hypothalamic–pituitary–adrenal axis by cytokines: actions and mechanisms of action. *Physiol. Rev.* **79,** 1–71 (1999).
35. Cannon, W. Organization for physiological homeostasis. *Physiol. Rev.* **9,** 399–431 (1929).
This classic theoretical paper and review describes the principles of homeostatic control.
36. Zong, W. X. & Thompson, C. B. Necrotic death as a cell fate. *Genes Dev.* **20,** 1–15 (2006).
37. Huang, H. & Tindall, D. J. Dynamic FoxO transcription factors. *J. Cell Sci.* **120,** 2479–2487 (2007).
38. Ghosh, S. & Karin, M. Missing pieces in the NF-κB puzzle. *Cell* **109,** S81–S96 (2002).
39. Zong, W. X., Ditsworth, D., Bauer, D. E., Wang, Z. Q. & Thompson, C. B. Alkylating DNA damage stimulates a regulated form of necrotic cell death. *Genes Dev.* **18,** 1272–1282 (2004).
40. Gordon, S. & Taylor, P. R. Monocyte and macrophage heterogeneity. *Nature Rev. Immunol.* **5,** 953–964 (2005).
41. Hume, D. A. The mononuclear phagocyte system. *Curr. Opin. Immunol.* **18,** 49–53 (2006).
42. Kanda, H. *et al.* MCP-1 contributes to macrophage infiltration into adipose tissue, insulin resistance, and hepatic steatosis in obesity. *J. Clin. Invest.* **116,** 1494–1505 (2006).
43. Ceradini, D. J. & Gurtner, G. C. Homing to hypoxia: HIF-1 as a mediator of progenitor cell recruitment to injured tissue. *Trends Cardiovasc. Med.* **15,** 57–63 (2005).
44. Mantovani, A. *et al.* The chemokine system in diverse forms of macrophage activation and polarization. *Trends Immunol.* **25,** 677–686 (2004).
45. Gordon, S. Alternative activation of macrophages. *Nature Rev. Immunol.* **3**, 23–35 (2003).
46. Condeelis, J. & Pollard, J. W. Macrophages: obligate partners for tumor cell migration, invasion, and metastasis. *Cell* **124,** 263–236 (2006).
47. Majno, G. & Joris, I. Apoptosis, oncosis, and necrosis. An overview of cell death. *Am. J. Pathol.* **146,** 3–15 (1995).
48. Fink, S. L. & Cookson, B. T. Apoptosis, pyroptosis, and necrosis: mechanistic description of dead and dying eukaryotic cells. *Infect. Immun.* **73,** 1907–1916 (2005).
49. Rothlin, C. V., Ghosh, S., Zuniga, E. I., Oldstone, M. B. & Lemke, G. TAM receptors are pleiotropic inhibitors of the innate immune response. *Cell* **131,** 1124–1136 (2007).
50. Henson, P. M. & Hume, D. A. Apoptotic cell removal in development and tissue homeostasis. *Trends Immunol.* **27,** 244–250 (2006).
51. Ravichandran, K. S. & Lorenz, U. Engulfment of apoptotic cells: signals for a good meal. *Nature Rev. Immunol.* **7,** 964–974 (2007).
52. Werner, S. & Grose, R. Regulation of wound healing by growth factors and cytokines. *Physiol. Rev.* **83,** 835–870 (2003).
53. Tedgui, A. & Mallat, Z. Cytokines in atherosclerosis: pathogenic and regulatory pathways. *Physiol. Rev.* **86,** 515–581 (2006).
54. Hotamisligil, G. S. Inflammation and metabolic disorders. *Nature* **444,** 860–867 (2006).
55. Karin, M., Lawrence, T. & Nizet, V. Innate immunity gone awry: linking microbial infections to chronic inflammation and cancer. *Cell* **124,** 823–835 (2006).
56. Stearns, S. & Koella, J. *Evolution in Health and Disease* (Oxford Univ. Press, 2008).
57. Gould, S. J. & Lewontin, R. C. The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proc. R Soc. Lond. B* **21,** 581–598 (1979).

Acknowledgements I apologize to the many authors whose work could not be cited directly because of space limitations. I thank I. Brodsky, T. Horng, A. Iwasaki, E. Kopp, N. Palm and D. Stetson for critical reading of the manuscript. R.M. is an investigator of the Howard Hughes Medical Institute.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The author declares no competing financial interests. Correspondence should be addressed to the author (ruslan.medzhitov@yale.edu).

NATURE|Vol 454|24 July 2008|doi:10.1038/nature07205

Cancer-related inflammation

Alberto Mantovani[1,2], Paola Allavena[1], Antonio Sica[3] & Frances Balkwill[4]

The mediators and cellular effectors of inflammation are important constituents of the local environment of tumours. In some types of cancer, inflammatory conditions are present before a malignant change occurs. Conversely, in other types of cancer, an oncogenic change induces an inflammatory microenvironment that promotes the development of tumours. Regardless of its origin, 'smouldering' inflammation in the tumour microenvironment has many tumour-promoting effects. It aids in the proliferation and survival of malignant cells, promotes angiogenesis and metastasis, subverts adaptive immune responses, and alters responses to hormones and chemotherapeutic agents. The molecular pathways of this cancer-related inflammation are now being unravelled, resulting in the identification of new target molecules that could lead to improved diagnosis and treatment.

Links between cancer and inflammation were first made in the nineteenth century, on the basis of observations that tumours often arose at sites of chronic inflammation and that inflammatory cells were present in biopsied samples from tumours[1]. The idea that these processes are connected was out of favour for more than a century, but there has been a recent resurgence in interest. Several lines of evidence[1–4] (Box 1) — based on a range of findings, from epidemiological studies of patients to molecular studies of genetically modified mice — have led to a general acceptance that inflammation and cancer are linked.

Epidemiological studies have shown that chronic inflammation predisposes individuals to various types of cancer. It is estimated that underlying infections and inflammatory responses are linked to 15–20% of all deaths from cancer worldwide[1]. There are many triggers of chronic inflammation that increase the risk of developing cancer. Such triggers include microbial infections (for example, infection with *Helicobacter pylori* is associated with gastric cancer and gastric mucosal lymphoma), autoimmune diseases (for example, inflammatory bowel disease is associated with colon cancer) and inflammatory conditions of unknown origin (for example, prostatitis is associated with prostate cancer). Accordingly, treatment with non-steroidal anti-inflammatory agents decreases the incidence of, and the mortality that results from, several tumour types[5–7].

Box 1 | The evidence that links cancer and inflammation

- Inflammatory diseases increase the risk of developing many types of cancer (including bladder, cervical, gastric, intestinal, oesophageal, ovarian, prostate and thyroid cancer).
- Non-steroidal anti-inflammatory drugs reduce the risk of developing certain cancers (such as colon and breast cancer) and reduce the mortality caused by these cancers.
- Signalling pathways involved in inflammation operate downstream of oncogenic mutations (such as mutations in the genes encoding RAS, MYC and RET).
- Inflammatory cells, chemokines and cytokines are present in the microenvironment of all tumours in experimental animal models and humans from the earliest stages of development.
- The targeting of inflammatory mediators (chemokines and cytokines, such as TNF-α and IL-1β), key transcription factors involved in inflammation (such as NF-κB and STAT3) or inflammatory cells decreases the incidence and spread of cancer.
- Adoptive transfer of inflammatory cells or overexpression of inflammatory cytokines promotes the development of tumours.

The hallmarks of cancer-related inflammation include the presence of inflammatory cells and inflammatory mediators (for example, chemokines, cytokines and prostaglandins) in tumour tissues, tissue remodelling and angiogenesis similar to that seen in chronic inflammatory responses, and tissue repair. These signs of 'smouldering' inflammation[2] are also present in tumours for which a firm causal relationship to inflammation has not been established (for example, breast tumours). Indeed, inflammatory cells and mediators are present in the microenvironment of most, if not all, tumours, irrespective of the trigger for development.

Studies of genetically modified mice, adoptive-transfer experiments in mice, and analyses of human tumours have allowed researchers to begin to unravel the molecular pathways that link inflammation and cancer. Here we review current knowledge of the molecular and cellular pathways that link inflammation and cancer, and we describe how these pathways suppress effective antitumour immunity during tumour progression. We also discuss how cancer-related inflammation affects many aspects of malignancy, including the proliferation and survival of malignant cells, angiogenesis (which is required for the survival of cells within tumours of a certain size), tumour metastasis, and tumour response to chemotherapeutic drugs and hormones.

Advances in understanding the genetic pathways involved in cancer have led to the development of a range of therapies that target malignant cells. Understanding the pathways involved in cancer-related inflammation could enable the development of synergistic therapies that target 'the other half of the tumour' — that is, the inflammatory components of the microenvironment. Preclinical and early clinical studies are now suggesting how this might be achieved. However, despite considerable progress, important questions remain unanswered, as we discuss in the final section of this Review.

Connecting inflammation and oncogenes

The connection between inflammation and cancer can be viewed as consisting of two pathways: an extrinsic pathway, driven by inflammatory conditions that increase cancer risk (such as inflammatory bowel disease); and an intrinsic pathway, driven by genetic alterations that

[1]Istituto Clinico Humanitas IRCCS, Via Manzoni 56, Rozzano, 20089 Milan, Italy. [2]Istituto di Patologia Generale, Universita degli Studi di Milano, 20133 Milan, Italy. [3]Fondazione Humanitas per la Ricerca, Via Manzoni 56, Rozzano, 20089 Milan, Italy. [4]Centre for Cancer & Inflammation, Institute of Cancer and the Cancer Research UK Clinical Centre, Barts and The London School of Medicine and Dentistry, London EC1M 6BQ, UK.

cause inflammation and neoplasia (such as oncogenes) (Fig. 1). The intrinsic pathway was uncovered when addressing why inflammatory cells and mediators are present in the microenvironment of most, if not all, tumours and therefore are present in cases for which there is no epidemiological basis for inflammation. This finding raised the question of whether the genetic events that cause neoplasia in these cases are responsible for generating an inflammatory environment. This question has been addressed only recently, by using preclinical and clinical settings in which various oncogenetic mechanisms can be assessed.

An example of this type of study involves protein tyrosine kinases, which are encoded by prototypical transforming oncogenes. A useful clinical setting in which to explore the connection between these oncogenes and an inflammatory microenvironment is human papillary thyroid carcinoma. Rearrangement of the chromosome on which the gene encoding the protein tyrosine kinase RET is located (also referred to as the *RET*/*PTC* rearrangement) is a frequent early event in the pathogenesis of papillary thyroid carcinoma and is a necessary and sufficient event for this cancer to develop. In an appropriate cellular context, which is provided by freshly isolated human thyrocytes, the activation of RET induces a transcriptional program that is similar to that which occurs during inflammation[8] (Fig. 2). The transcriptome of cells in which RET has been activated includes messenger RNA encoding various factors: colony-stimulating factors (CSFs), which promote the survival of leukocytes and their recruitment from the blood to the tissues; interleukin 1β (IL-1β), one of the main inflammatory cytokines; cyclooxygenase 2 (COX2), which is frequently expressed by cancerous cells and is involved in the synthesis of prostaglandins; chemokines that can attract monocytes and dendritic cells (CC-chemokine ligand 2 (CCL2) and CCL20); chemokines that promote angiogenesis (such as IL-8; also known as CXC-chemokine ligand 8 (CXCL8)); the chemokine receptor CXC-chemokine receptor 4 (CXCR4), which binds to CXCL12; extracellular-matrix-degrading enzymes; and the adhesion molecule lymphocyte selectin (L-selectin). Key protein components of the RET-activated 'inflammatory' program were found in tumour specimens taken by biospy, and larger amounts of these inflammatory molecules were found in the primary tumours of patients with lymph-node metastasis than in primary tumours in the absence of lymph-node metastasis[8,9]. These results show that an early genetic event that is necessary and sufficient for the development of a human tumour directly promotes the build-up of an inflammatory microenvironment[8]. Although these[8,9] and other results[10] connect the activation of protein-tyrosine-kinase-encoding oncogenes to inflammation, the precise roles of the various components of the inflammatory microenvironment in the progression of tumours remain to be defined.

Members of the *RAS* family are the most frequently mutated dominant oncogenes in human cancer, and activated oncogenic components of the RAS–RAF signalling pathway, in turn, induce the production of tumour-promoting inflammatory chemokines and cytokines[11–13]. Another oncogene, *MYC*, encodes a transcription factor that is overexpressed in many human tumours; deregulated expression of this gene initiates and maintains key aspects of the tumour phenotype. In addition to promoting cell-autonomous proliferation, MYC instructs the remodelling of the extracellular microenvironment, with inflammatory cells and mediators having important roles in this process. In a mouse model of MYC-dependent cancer of the β-cells (which are insulin-producing pancreatic islet cells), the first wave of angiogenesis results from MYC-induced production of the inflammatory cytokine IL-1β[14]. The MYC-activated transcriptional program also elicits the production of several chemokines that recruit mast cells. Mast cells have long been known to drive angiogenesis, and in this case (following IL-1β) they sustain new blood-vessel formation and tumour growth[15].

These studies of RAS family members and MYC show that dominant oncogenes promote the formation of a tumour-promoting tissue microenvironment (the intrinsic pathway), but the findings do not address the issue of the interplay between oncogenes and inflammatory conditions that increase the risk of developing cancer (the extrinsic pathway) (Fig. 1). This interplay is likely to occur in pancreatic carcinoma, in which both pancreatitis and mutations in the gene encoding K-RAS

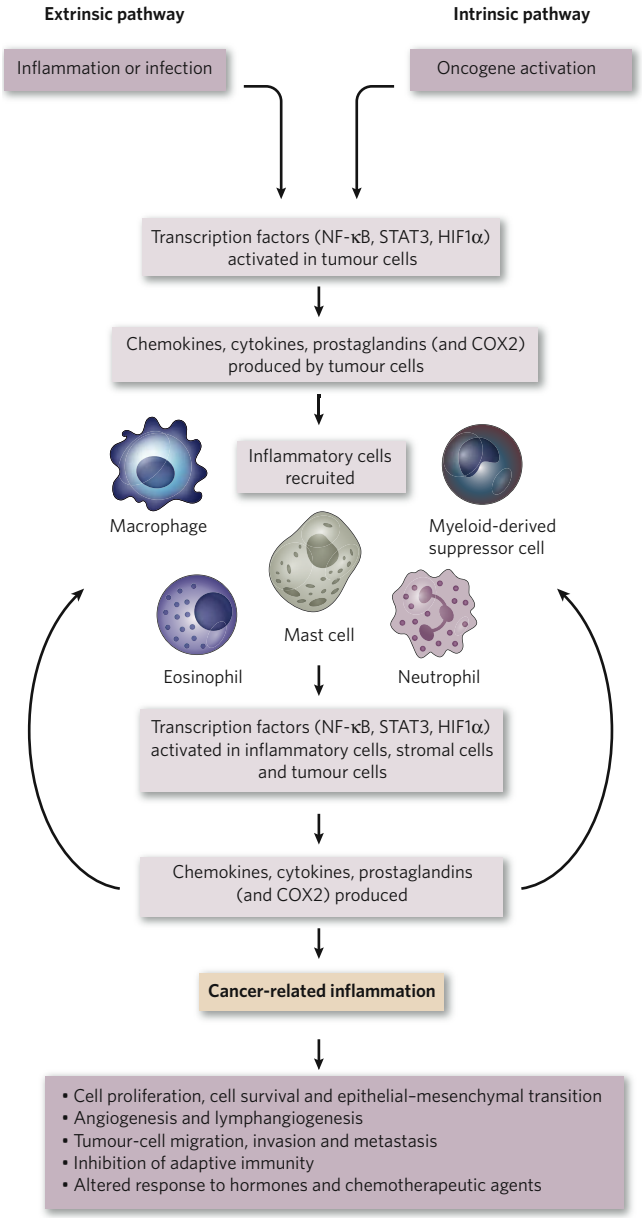


Figure 1 | Pathways that connect inflammation and cancer. Cancer and inflammation are connected by two pathways: the intrinsic pathway and the extrinsic pathway. The intrinsic pathway is activated by genetic events that cause neoplasia. These events include the activation of various types of oncogene by mutation, chromosomal rearrangement or amplification, and the inactivation of tumour-suppressor genes. Cells that are transformed in this manner produce inflammatory mediators, thereby generating an inflammatory microenvironment in tumours for which there is no underlying inflammatory condition (for example, breast tumours). By contrast, in the extrinsic pathway, inflammatory or infectious conditions augment the risk of developing cancer at certain anatomical sites (for example, the colon, prostate and pancreas). The two pathways converge, resulting in the activation of transcription factors, mainly nuclear factor-κB (NF-κB), signal transducer and activator of transcription 3 (STAT3) and hypoxia-inducible factor 1α (HIF1α), in tumour cells. These transcription factors coordinate the production of inflammatory mediators, including cytokines and chemokines, as well as the production of cyclooxygenase 2 (COX2) (which, in turn, results in the production of prostaglandins). These factors recruit and activate various leukocytes, most notably cells of the myelomonocytic lineage. The cytokines activate the same key transcription factors in inflammatory cells, stromal cells and tumour cells, resulting in more inflammatory mediators being produced and a cancer-related inflammatory microenvironment being generated. Smouldering cancer-related inflammation has many tumour-promoting effects.

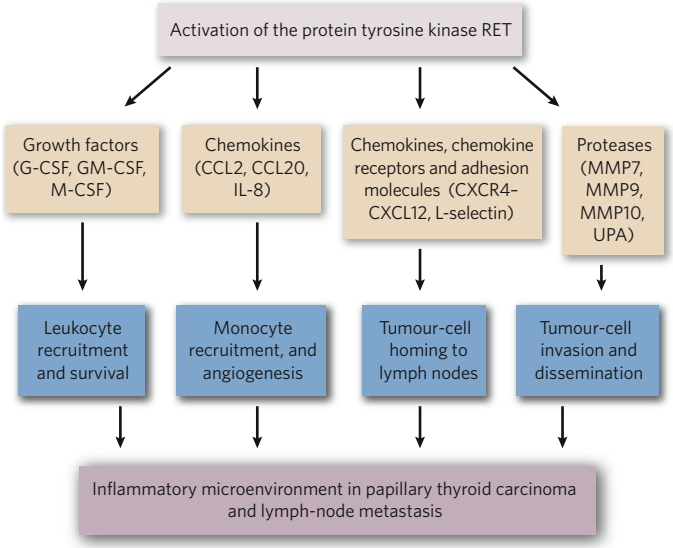


Figure 2 | Oncogenes and cancer-related inflammation. A class of oncogenes encodes protein tyrosine kinases that are persistently activated in a ligand-independent manner as a result of mutation or chromosomal rearrangement. RET is representative of these activated oncogenes. A chromosomal rearrangement that affects *RET* is a frequent early event in the pathogenesis of human papillary thyroid carcinoma. In human thyrocytes maintained in short-term culture, RET activates an inflammatory transcriptional program, the components of which are found in tumours obtained from patients. The inflammatory mediators that are produced in response to this program, together with their effects on tumour cells and inflammatory cells, are indicated. CCL, CC-chemokine ligand; CSF, colony-stimulating factor; CXCL12, CXC-chemokine ligand 12; CXCR4, CXC-chemokine receptor 4; IL-8, interleukin 8; L-selectin, lymphocyte selectin; MMP, matrix metalloproteinase; UPA, urokinase-type plasminogen activator.

are frequently found. In a relevant mouse model, adult mice are resistant to mutated *Kras*-induced pancreatic carcinogenesis[11]. Both mild chronic pancreatitis (possibly mirroring the clinical epidemiology) and mutated *Kras* are required to induce pancreatic intra-epithelial neoplasia and invasive ductal carcinoma[11]. Thus, although the RAS–RAF pathway[12,13] can drive tumour-promoting inflammation to a certain extent, an extrinsic inflammatory condition (pancreatitis) is needed to drive carcinogenesis in mice and presumably in humans.

Tumour-suppressor proteins can also regulate the production of inflammatory mediators. Examples of such proteins are von Hippel-Lindau tumour suppressor (VHL), transforming growth factor-β (TGF-β) and phosphatase and tensin homologue (PTEN)[16–20]. VHL is a component of a molecular complex that targets the transcription factor hypoxia-inducible factor 1α (HIF1α) for degradation. HIF1α promotes the cellular and tissue response to hypoxia, including angiogenesis. It also functionally interacts with the transcription factor nuclear factor-κB (NF-κB) (discussed later), resulting in the production of the major inflammatory cytokine tumour-necrosis factor-α (TNF-α) and the chemokine receptor CXCR4 in human renal-cell carcinoma cells, as well as in other malignant cell types[2,16,19,20]. The production of CXCR4 is particularly relevant, because *CXCR4* expression is frequently upregulated in human cancer and CXCR4 is involved in metastasis[16] (discussed later).

Recent evidence from a mouse model of cancer links TGF-β, a tumour-suppressor protein that is frequently involved in the progression of human cancer, to tumour-promoting inflammation[21]. In an animal model of breast carcinoma, inactivation of the gene encoding the type II TGF-β receptor (which initiates carcinogenesis by preventing the actions of TGF-β) unleashes the production of CXCL5 and CXCL12. These chemokines attract cells known as myeloid-derived suppressor cells (MDSCs), which belong to the myelomonocytic lineage. MDSCs are potent suppressors of the adaptive immune response to tumours and directly facilitate metastasis. It will be important to assess whether this pathway occurs in human tumours in which the TGF-β receptor is involved.

Thus, the various types of oncogene (such as those encoding protein tyrosine kinases, RAS and RAF, transcription factors and tumour-suppressor proteins), irrespective of their molecular class or mode of action, all coordinate inflammatory transcriptional programs. And these oncogene-coordinated inflammatory responses seem to have aspects in common: a link to angiogenesis, and the recruitment of cells of myelomonocytic origin. Several issues remain to be fully elucidated, including which components of inflammation are essential and which are redundant, the relative importance of these components in carcinogenesis in different tissues, and the relevance of these components to different types of cancer in humans.

Key factors in cancer-related inflammation

In the panoply of molecules involved in cancer-related inflammation, key endogenous (intrinsic) factors can be identified. These include transcription factors (such as NF-κB and signal transducer and activator of transcription 3 (STAT3)) and major inflammatory cytokines (such as IL-1β, IL-6, IL-23 and TNF-α)[4,22–26]. The main inflammatory cytokines were discussed earlier, so this section focuses on transcription factors.

NF-κB is a key coordinator of innate immunity and inflammation, and has emerged as an important endogenous tumour promoter[4]. NF-κB is crucial both in the context of tumour or potential tumour cells and in the context of inflammatory cells. In these cell types, NF-κB operates downstream of the sensing of microorganisms or tissue damage by the Toll-like receptor (TLR)–MyD88 signalling pathway, and by signalling pathways mediated by the inflammatory cytokines TNF-α and IL-1β. In addition, NF-κB can be activated as a result of cell-autonomous genetic alterations (amplification, mutations or deletions)[27] in tumour cells.

In tumour cells and epithelial cells at risk of transformation by carcinogens, as well as in inflammatory cells, NF-κB activates the expression of genes encoding inflammatory cytokines, adhesion molecules, enzymes in the prostaglandin-synthesis pathway (such as COX2), inducible nitric oxide synthase (iNOS; also known as NOS2) and angiogenic factors. In addition, one of the important functions of NF-κB in tumour cells or cells targeted by carcinogenic agents is promoting cell survival, by inducing the expression of anti-apoptotic genes (such as *BCL2*). There is also accumulating evidence of interconnections and compensatory pathways between the NF-κB and HIF1α systems[28–30], linking innate immunity to the response to hypoxia.

There is unequivocal evidence that NF-κB is involved in tumour initiation and progression in tissues in which cancer-related inflammation typically occurs (such as the gastrointestinal tract and the liver)[31,32]. This evidence is based on various genetic studies, such as tissue-specific targeting of genes that encode components of the IκB kinase (IKK) complex. This complex phosphorylates inhibitor of NF-κB (IκB), causing it to dissociate from NF-κB and allowing NF-κB to translocate to the nucleus, where it can exert its function as a transcription factor. It should be noted that genetic targeting of NF-κB in liver epithelial cells can have divergent effects in different models of carcinogenesis, possibly depending on the balance between promoting apoptosis that has already been initiated and triggering compensatory cell proliferation[32,33].

The NF-κB pathway is tightly controlled by inhibitors that function at various stages of the pathway. An example is TIR8 (also known as SIGIRR), a member of the IL-1-receptor family. TIR8 has a single immunoglobulin domain, a long cytoplasmic tail, and a Toll/IL-1 receptor (TIR) domain that differs from that of other members of the IL-1-receptor family. It inhibits signalling through TLRs and the IL-1 receptor and is highly expressed in the intestinal mucosa. Deficiency in the gene that encodes TIR8 is associated with increased susceptibility to intestinal inflammation and carcinogenesis[34,35]. Thus, the balance of inhibitors and activators tunes the extent to which the NF-κB pathway operates as an endogenous tumour promoter.

Support for the connection between cancer and inflammation is further strengthened by studies of the role of NF-κB in tumour-infiltrating leukocytes. For example, by using the strategy mentioned above,

myeloid-lineage-specific inactivation of the gene encoding IKK-β was found to inhibit cancer-related inflammation in the intestine, as well as colitis-associated cancer, unequivocally showing that inflammatory cells are involved in carcinogenesis in this tissue[31]. In established, advanced tumours, which typically have a microenvironment of smouldering inflammation[2], tumour-associated macrophages (TAMs) have delayed and defective NF-κB activation[36]. Evidence suggests that homodimers of the p50 subunit of NF-κB (a negative regulator of the NF-κB pathway) are responsible for this sluggish activation of NF-κB in TAMs and for the protumour phenotype of these cells[37]. Thus, NF-κB seems to function as a 'rheostat' whose function can be tuned to different levels, a property that enables the extent of inflammation to be regulated. Such regulation allows the vigorous inflammation (for example, in inflammatory bowel disease) that predisposes individuals towards developing cancer to be sustained, and enables TAMs to sustain the smouldering inflammatory microenvironment present in established metastatic neoplasia.

Similar to NF-κB, STAT3 is a point of convergence for numerous oncogenic signalling pathways[22]. This transcription factor is constitutively activated both in tumour cells and in immune cells, and is involved in oncogenesis and inhibition of apoptosis[38]. The activation of STAT3 in tumour cells has also been shown to increase the capacity of tumours to evade the immune system, by inhibiting the maturation of dendritic cells[39] and suppressing the immune response[40].

Tumour-infiltrating leukocytes

A leukocyte infiltrate, varying in size, composition and distribution, is present in most, if not all, tumours. Its components include TAMs and related cell types, mast cells and T cells. There is evidence (based on adoptive-transfer studies, cell-depletion studies, clinical correlations and gene-manipulation studies) that each of these bone-marrow-derived components can be involved in carcinogenesis and/or tumour invasion and metastasis[41–44].

In this section, we focus on TAMs. TAMs are an important component of the leukocyte infiltrate, and studies of TAMs formed the basis for the model that leukocyte infiltrates are involved in tumour progression. Plasticity and diversity are hallmarks of mononuclear phagocytes. In addition to conventional TAMs, related cell populations (for example, a TIE2-expressing monocyte subset, MDSCs and myeloid dendritic cells) have been linked to a protumour inflammatory microenvironment[45,46]. The ontogenetic relationship between these cell types and their relative importance in the context of tumours remain to be elucidated.

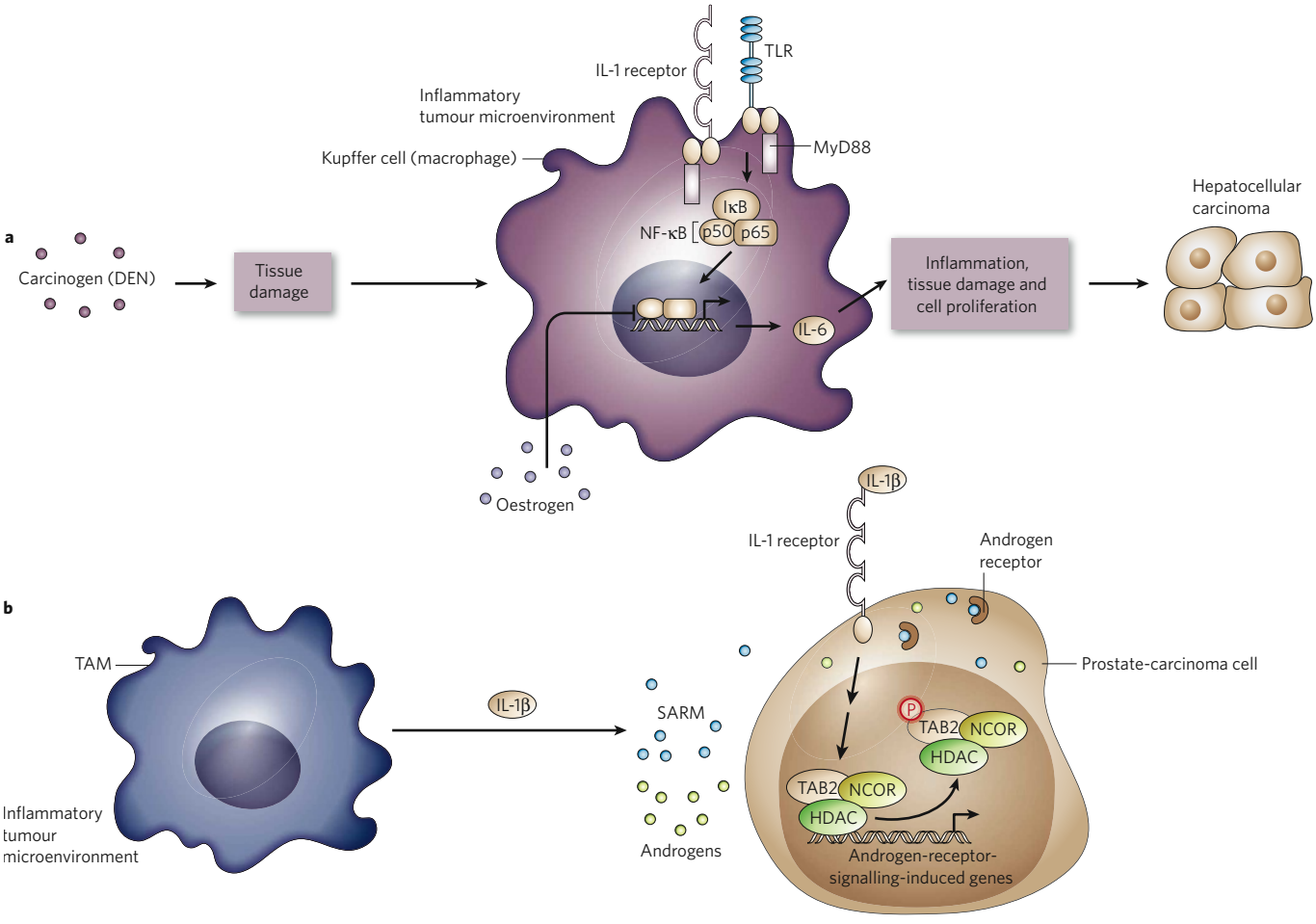


Figure 3 | Hormones and inflammation. Hormones and inflammation are each involved in a classic pathway that promotes tumour development. **a**, Liver. The carcinogen diethylnitrosamine (DEN) can activate the MyD88 signalling pathway in Kupffer cells (a type of macrophage) in the liver (presumably through TLRs or IL-1 receptors), resulting in the production of IL-6 (ref. 61). IL-6, in turn, promotes inflammation, tissue damage, compensatory cell proliferation and, ultimately, formation of a liver tumour. Oestrogens interfere with the activity of NF-κB, the transcription factor that regulates IL-6 production, and thereby protect against carcinogenesis[61]. **b**, Prostate. Patients with prostate tumours are often treated with drugs known as selective androgen-receptor modulators (SARMs), which reduce the growth-promoting effects of male sex hormones (androgens) on the tumour. TAMs can, however, produce IL-1β, which results in the desired action of SARMs being reversed, from inhibitors of androgen-receptor-signalling-induced gene expression (as intended) to activators. This process involves TAB2 (TGF-β-activated kinase 1 (TAK1)-binding protein 2), a sensor of inflammatory signals. TAB2 is a component of a co-repressor complex that includes the nuclear-receptor co-repressor (NCOR) and a histone deacetylase (HDAC). IL-1β-mediated signalling results in the phosphorylation of TAB2, thereby lifting the repression of transcription. SARMs thus promote tumour-cell proliferation instead of inhibiting it. (Figure modified, with permission, from ref. 92.)

TAMs and related cell types in mouse and human tumours generally have an M2 phenotype, which is oriented towards promoting tumour growth, remodelling tissues, promoting angiogenesis and suppressing adaptive immunity[45,47]. Signals that are derived from regulatory T cells present in tumours or from the tumour cells themselves (including macrophage CSF (M-CSF), IL-10 and TGF-β)[45,47,48] might account for this polarization to the M2 phenotype of macrophages that have been recruited into tumours. However, a stringent analysis of the molecular mechanisms responsible for this functional polarization during tumour progression has not been carried out.

Leukocyte infiltration is also interconnected with angiogenesis, which is required in tumours of a certain size. The pro-angiogenic protein vascular endothelial growth factor (VEGF) and related molecules are potent monocyte attractants and contribute to the recruitment of monocytes into primary tumours and the metastatic niche[46,49–51]. In turn, the recruited leukocytes provide an indirect (VEGF-independent) pathway of angiogenesis, through the secretion of pro-angiogenic factors[41,51–55]. It is possible therefore that the inhibition of leukocyte recruitment will improve the activity of current anti-angiogenic therapies for patients with cancer.

Cancer-related inflammation and adaptive immunity

There is strong evidence from genetic studies of mouse models that cells of the adaptive immune system carry out surveillance and can eliminate nascent tumours (a process called immuno-editing)[56]. Innate immune responses, which manifest as inflammation, are crucial for the initiation of adaptive immune responses. Therefore, the seemingly divergent effects of inflammation and immuno-editing are paradoxical. Yet, a recent study in mice shows that the TLR adaptor MyD88 (which is involved in innate immune responses) has a key role in promoting tumour development and that inflammation-induced carcinogenesis and immuno-editing can occur in the same tumour model[57].

Another aspect of the complex interplay between adaptive immunity and cancer-related inflammation was shown by studies in a mouse model of cancer caused by human papilloma virus. In this system, antibodies are deposited in the tumour stroma. These antibodies then function as a 'remote-control system', binding to unidentified molecules in the extracellular matrix and thereby activating inflammatory responses that promote cancer progression[58].

In clinically overt neoplasia, effective adaptive immune responses are suppressed through the activation of several pathways. For example, the differentiation and activation of dendritic cells, which are the key initiators of adaptive immune responses, are inhibited by signals (such as IL-10) present in the tumour microenvironment. In addition, tumours are frequently infiltrated by regulatory T cells, which suppress both adaptive and innate immune responses. And MDSCs proliferate in tumour-bearing hosts; these cells, as well as conventional TAMs, are potent suppressors of antitumour immunity[46,47]. Thus, in cancer-related inflammation, multiple pathways are set in motion in to suppress effective antitumour immunity in established tumours. The interplay between these pathways, their hierarchy, and whether they can be targeted for therapy remain largely to be determined.

Cancer-related inflammation and sex hormones

Sex steroid hormones mediate a classic, clinically relevant pathway of tumour promotion in breast and prostate cancer and have been a target for therapeutic intervention since George Beatson's discovery of hormone-dependent breast cancer at the end of the nineteenth century[59].

Recent studies, however, have uncovered an unexpected relationship between sex steroid hormones and cancer (Fig. 3). For carcinoma of the prostate, which is an androgen-dependent tumour, sensitivity to stimulation with hormones is regulated by selective androgen-receptor modulators. The inflammatory cytokine IL-1β, which is produced by macrophages in the tumour microenvironment, converts these receptor modulators from being inhibitory to stimulatory[60]. It has long been known that females are less susceptible to cancer at sites, such as the liver, that are not conventional target organs of sex steroid hormones. After studies of a mouse model of liver carcinogenesis, Willscott Naugler *et al.*[61] recently reported that the sex difference in tumour susceptibility resulted from a downregulation of IL-6 production by macrophages in response to oestrogens. In addition, in male mice, IL-6 production was triggered to a much greater extent in response to carcinogen-mediated tissue damage (which induces IL-6 production by activating MyD88-dependent TLR- and/or IL-1-receptor signalling pathways). Thus, connections are emerging between the two classic tumour-promoting pathways — inflammation and sex steroid hormones — bringing together the pioneering efforts of Beatson and Rudolf Virchow in studying inflammatory cells in tumours.

Inflammatory pathways in invasion and metastasis

Most studies of the mechanisms of cancer-related inflammation have focused on the early stages of cancer, but inflammatory mediators and cells are also involved in the migration, invasion and metastasis of malignant cells.

Chemokine receptors and their ligands direct the movement of cells during inflammation, cancer and the maintenance of tissue homeostasis, by affecting cell motility, invasiveness and survival[16]. On transformation, many cells start to express chemokine receptors and thereby use chemokines to aid in their migration to, and survival at, sites that are distant from the original tumour[16,62,63]. For example, the chemokine receptor CXCR4 and its ligand CXCL12 are important for cell movement in both homeostatic and disease states[63]. CXCR4 is frequently expressed by malignant cells[16], and the amount of CXCR4 expressed by primary human tumours correlates with the extent to which metastasis to the lymph nodes occurs in colorectal, breast, liver and oesophageal cancer[64–66]. Other functional chemokine receptors (including CX_3C-chemokine receptor 1 (CX_3CR1), CC-chemokine receptor 1 (CCR1), CCR7, CCR9, CCR10, CXCR1, CXCR2, CXCR3, CXCR5 and CXCR7) are also expressed by malignant cells from a variety of tissues and are implicated in organ-specific metastasis[67–72]; for example, the expression of CCR7 correlates with lymph-node metastasis, and expression of CCR9 with metastasis of melanoma to the small intestine. Malignant melanoma cells express many of the above receptors, perhaps explaining why melanomas are highly metastatic.

How do malignant cells acquire the ability to express chemokine receptors? Several mechanisms have been defined. Autocrine and paracrine extracellular signals, as well as genetic and epigenetic alterations, might each contribute to this change. For example, the previously discussed mutation in *VHL* and the chromosomal rearrangement that affects *RET* induce the expression of CXCR4 on initiated cells. Regardless of the mechanism, it is clear that acquisition of chemokine-receptor expression is a common attribute of malignant cells of epithelial or mesenchymal origin that do not normally express these receptors and that this can occur even at the early stages of malignancy[65].

The invasive capacity of malignant cells can increase in the presence of inflammatory cytokines such as TNF-α, IL-1β and IL-6 (ref. 2), possibly as a result of the upregulation of chemokine-receptor expression elicited by these cytokines[73]. For example, autocrine TNF-α-mediated signalling upregulates the expression of functional CXCR4 by ovarian cancer cells[74], and stable knockdown of mRNA encoding this cytokine reduces the expression of both CXCR4 and its ligand CXCL12 by the malignant cells, inhibiting colonization of the peritoneal cavity, angiogenesis and spread to sites distant from the peritoneal cavity[73]. TNF-α is also a potent stimulator of epithelial–mesenchymal transition by breast cancer cells[75], as is activation of NF-κB signalling.

A further link between NF-κB signalling and metastasis was found by studying a genetic model of prostate cancer in mice, in which inactivation of the gene encoding a major component of the NF-κB signalling pathway, IKK-α, was found to reduce metastatic spread[76]. The mechanism by which this occurs was found to involve activation of receptor activator of NF-κB (RANK) at the surface of malignant prostate epithelial cells in a paracrine manner, by RANK ligand derived from the leukocytes infiltrating the primary tumours. IKK-α, which is involved in the RANK signalling pathway, then inhibited expression of the metastasis-suppressor protein maspin, hence promoting the metastatic phenotype. Therefore, removing IKK-α had the opposite effect. It would be interesting to determine whether

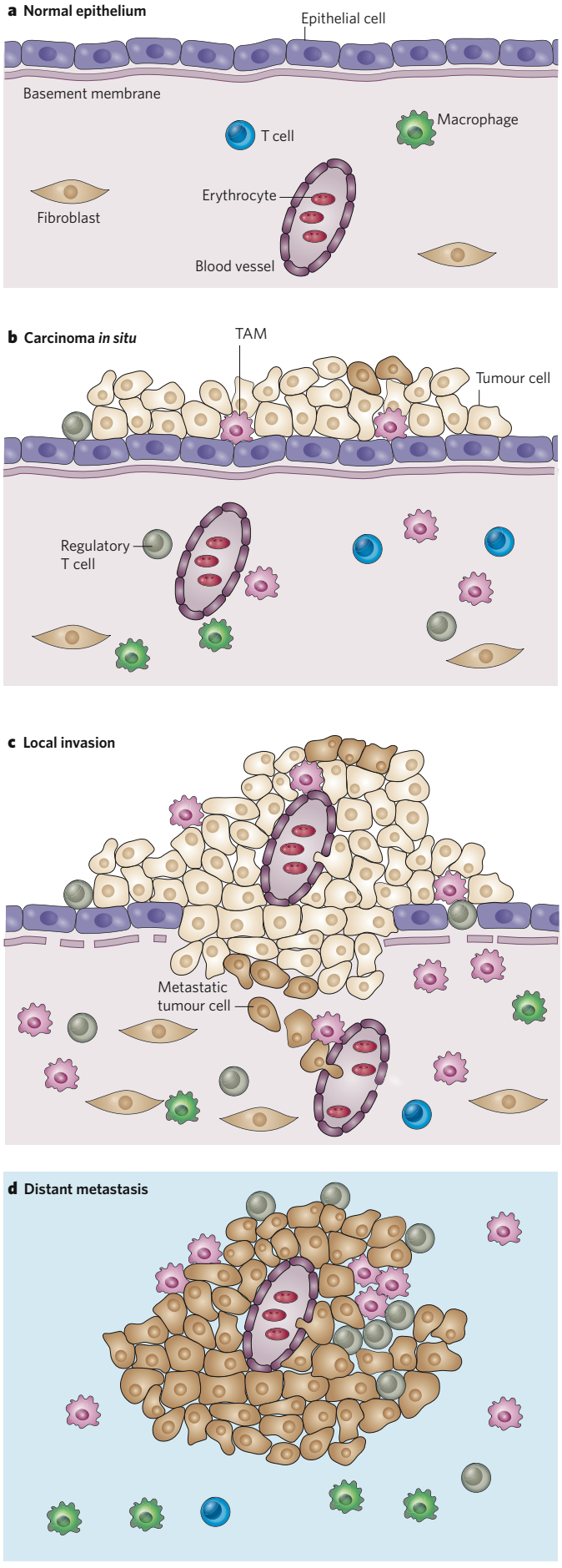


Figure 4 | Inflammation and the malignant progression of epithelial tumours. a, Normal epithelium. The proliferation of epithelial cells and the homeostatic trafficking of leukocytes in the epithelium is regulated by autocrine and paracrine chemokine- and cytokine-mediated signalling. **b**, Carcinoma *in situ*. The extrinsic (inflammatory) pathway and the intrinsic (oncogenic) pathway induce the production of chemokines and cytokines. These factors attract a tumour-promoting infiltrate (which contains, for example, TAMs and regulatory T cells, as shown here), and they also promote angiogenesis. The expression of chemokine receptors can be induced on initiated cells (that is, early tumour cells). These receptors then aid in tumour-cell survival and might be necessary (but are not sufficient) for invasion across the basement membrane. Both autocrine and paracrine networks of cytokines and chemokines are involved in these processes. **c**, Local invasion. The chemokines and cytokines continue to attract and modulate a tumour-promoting infiltrate. They also promote angiogenesis and control tissue remodelling (for example, changes in the basement membrane). Operating in paracrine and autocrine loops, these factors induce the expression of genes associated with survival, invasion and migration in cells that have enough oncogenic changes to allow them to invade the basement membrane. The chemokines and cytokines are also involved in the intravasation of tumour cells into blood vessels and in lymphatic spread. **d**, Distant metastasis. The autocrine and paracrine chemokine- and cytokine-mediated signalling promotes the survival of malignant cells in distant organs, again attracting a tumour-promoting infiltrate and stimulating angiogenesis.

chemokine ligands and their receptors also contribute to the effects of IKK-α and whether IKK-α is implicated in other metastatic pathways.

Other cells within the tumour also affect processes in the later stages of cancer. TAMs have been described as "obligate partners for tumour-cell migration, invasion and metastasis"[77]. The first experiments to show this conclusively involved a genetic model of breast cancer in macrophage-deficient mice[44]. The tumours developed normally but were unable to form pulmonary metastases in the absence of macrophages. The mechanism by which metastasis occurs in this case involves a paracrine loop of tumour-cell M-CSF and macrophage epidermal growth factor[44], with intravasation assisted by direct interactions between tumour cells and TAMs[78].

In addition, inflammatory macrophages increase peritoneal dissemination of tumour cells and metastatic spread in an ovarian cancer model[79]. The ability of macrophages to aid in ovarian tumour-cell migration and invasion can also be modelled *in vitro*[48,80]. Co-culture of macrophages with tumour cells was shown to increase their invasive capacity in an NF-κB-dependent and TNF-α-dependent manner.

In summary, chemokines and cytokines coordinate autocrine and paracrine interactions between malignant cells and infiltrating leukocytes. These interactions increase the migration, invasion and survival of malignant cells. They also affect the growth of the primary tumour and the ability of tumour cells to colonize the metastatic niche[16,63,67,81] (Fig. 4).

Cancer-inhibitory inflammation

Although numerous experimental and clinical results point to inflammation having protumour activity, some evidence does not fit into this general pattern. For example, a marked chronic inflammatory response such as that in psoriasis is not associated with an increased risk of developing skin cancer[82]. Also, in certain tumours or subsets of tumours, the presence of inflammatory cells is associated with better prognosis (for example, eosinophils in colon tumours, and TAMs in a subset of breast tumours and pancreatic tumours). These observations are likely to reflect that inflammatory cells can destroy tumour cells, in addition to normal tissue cells[41]. For example, appropriately activated macrophages, a prototypical component of cancer-related inflammation, can kill tumour cells and elicit cancer-destructive inflammatory responses centred on the blood-vessel wall, although in most cases their tumour-promoting properties prevail[41]. Evidence indicates that NF-κB is important in determining this balance between the protumour and antitumour properties of macrophages[37,83], thus NF-κB could be targeted to 're-educate' tumour-promoting macrophages towards an antitumour function.

Box 2 | Unanswered questions about cancer-related inflammation

1. Is inflammation sufficient for cancer development?
2. Despite the diversity of tumours and oncogenic pathways, are there aspects of cancer-related inflammation that are common to all malignancies?
3. How can the balance between 'bad' inflammation and 'good' inflammation be altered to favour adaptive immunity instead of tumour development?
4. What is the relationship between MDSCs and TAMs?
5. What is the clinical relevance of the connections between sex steroid hormones and inflammation?
6. What is the best way to target cancer-related inflammation in patients with cancer? This is the most difficult question.

The importance of this balance is evident in psoriasis. Psoriasis is a T helper 1 (T_H1)-cell-mediated disease that involves a massive accumulation of neutrophils and monocytes in the skin, the latter of which are likely to become macrophages with an M1 phenotype, which have antitumour activity. Hence, the type of inflammation found in psoriasis does not promote the development of skin cancers because of the presence of macrophages that can destroy any nascent tumour cells. The dual potential of cancer-related inflammation (cancer promoting versus cancer inhibiting) may also be affected by the tissue type. In a skin tumour model, the overexpression of NF-κB was found to inhibit invasive epidermal neoplasia[84], whereas blocking NF-κB activity inhibited the development of experimental liver and colon cancers[31,32].

The concept that the activation of innate immune responses can promote a protective response to cancer is not new. In the late nineteenth century, William Coley noted that some patients with cancer who had severe postoperative infections at the tumour site underwent spontaneous and sustained tumour regression[85]. He then developed Coley's mixed toxins, a filtrate from cultures of *Streptococcus pyogenes* and *Serratia marcescens*, which was administered into the tumour or the surrounding tissues in patients with a range of advanced cancers. Although both the technique and the results were controversial, even at the time, Coley documented cases of the long-term survival of individuals with malignancies that remain a major challenge to treat now. And even before Coley's time, there was evidence for the regression of cancer after certain bacterial infections. More recently, this approach was adapted successfully to treat patients with bladder cancer by administering *Mycobacterium bovis* bacillus Calmette–Guérin (BCG). Such treatments probably trigger a 'good' inflammatory response (through TLRs) that not only promotes the differentiation of monocytes into macrophages with an M1 phenotype but also promotes the development of a sustained and effective adaptive immune response to the tumour. This type of response might also contribute to successful chemotherapy or radiotherapy, according to recent data obtained by Lionel Apetoh and colleagues[86]. After treating experimentally induced breast cancers, the authors found that dying tumour cells were able to cross-present antigen to dendritic cells in a TLR4- and MyD88-dependent manner, as well as trigger protective immune responses through a 'danger signal' (HMGB1), again by signalling through TLR4. However, when tumours were grown in mice with a mutant TLR4, the efficacy of chemotherapy and radiotherapy was reduced. Moreover, patients with breast cancer who have a mutation in TLR4 were found to have a higher frequency of metastasis.

The exact mechanisms by which a 'good' immune response can be reliably triggered during anticancer therapy are not entirely clear. It will be important to find the optimal stimuli to change a tumour-promoting (T_H2 cell and M2 macrophage) microenvironment to a tumour-inhibiting (T_H1 cell and M1 macrophage) microenvironment, and to understand the signalling mechanisms involved.

The big questions

The connection between inflammation and cancer is now generally accepted, but several questions remain. Some of these outstanding questions are listed in Box 2 and discussed in detail in this section.

First, it is unclear whether inflammation is sufficient for the development of cancer. That is, can inflammation cause neoplasia in the absence of an exogenous carcinogenic agent? Several lines of evidence provide hints that it can. In a mouse model of bowel inflammation caused by IL-10 deficiency, the frequency of DNA mutations observed in the colon in the absence of exogenous carcinogens was 4–5-fold greater than in IL-10-sufficient mice[87]. In addition, a comparison of human tumours and the appropriate normal tissues showed a higher frequency of random mutations in tumour cells[88]. The only exception to the exceedingly low frequency of random mutations found in normal cells ($< 1 \times 10^{-8}$ per base pair) was an inflamed tissue[88]. One candidate for an endogenous 'inflammatory carcinogen' is reactive oxygen species. Neutrophils, for example, have been shown to inhibit base-excision repair in an alveolar epithelial cell line, an effect that is mediated by a product of myeloperoxidase activity, hypochlorous acid[89]. Interestingly, a polymorphism in the promoter of the gene encoding myeloperoxidase has been associated with resistance to the development of lung cancer in smokers[90]. Therefore, although the evidence[87–91] suggests that inflammation causes cancer, formal proof is still required.

The second issue is that diversity and plasticity are characteristics of chronic inflammation and its main orchestrator, macrophages. Studies using classic histological techniques and a variety of model systems show that cancer-related inflammation differs between tumour types. It will be important to define which cellular and molecular components are common to all cancer-promoting inflammatory responses, and which are specific to particular tissues and tumour types.

The third issue relates to tipping the balance between cancer-promoting inflammatory responses and cancer-inhibiting inflammatory responses. A key point that needs to be addressed is how to activate an appropriate antitumour adaptive immune response.

The fourth open question relates to MDSCs, which have recently emerged as an important factor in cancer-related inflammation[46]. The definition of MDSCs is operational and includes a heterogeneous set of cells in the peripheral blood and spleen. Are these cells a distinct population, or do they belong to a continuum of TAM differentiation?

Fifth, the emergence of a connection between sex steroid hormones, inflammation and cancer is a major conceptual advance that provides the first link between these two classic tumour-promoting pathways[92]. This connection between inflammation and sex steroid hormones probably reflects the interplay between inflammation and hormones during reproduction. Analysing this link in human cancer and its implications for resistance to hormonal therapies could have a huge impact in the clinic, given the widespread use of selective androgen-receptor modulators and selective oestrogen-receptor modulators (for example, tamoxifen) to inhibit the growth-stimulating effects of sex hormones on prostate cancer and breast cancer, respectively.

Last, the biggest question is whether knowledge about cancer-related inflammation can be translated into useful approaches to preventing, diagnosing and treating cancer. Malignant cells are 'moving targets' that can become resistant to even the most sophisticated targeted drugs. Using a combination therapy that attacks both malignant cells and the 'other half' of the tumour mass (that is, the inflammatory cells) could be more effective and might elicit long-lasting adaptive immunity to the transformed cells.

Many drugs that could target cancer-related inflammation — for example, chemokine-receptor antagonists and cytokine-receptor antagonists, and COX inhibitors — are in clinical trials for other diseases. In terms of cancer, phase I/II trials of antagonists of IL-6, the IL-6 receptor, CCL2, CCR4 and CXCR4 are underway for a range of epithelial and haematopoietic malignancies. The first (phase I/II) clinical trials of TNF-α antagonists in patients with advanced cancer have resulted in disease stabilization and some partial responses[93–95], particularly for those with renal-cell carcinoma[95]. Also, a structural analogue of thalidomide, lenalidomide, that inhibits production of several inflammatory cytokines has been shown to be active against advanced myeloma when combined with dexamethasone[96]. In addition, COX2 inhibitors have been shown to prevent the recurrence of both sporadic adenomatous

polyps and adenomas in people with a genetic predisposition to developing these[97,98].

Drugs that target cancer-related inflammation have the potential to re-educate a tumour-promoting inflammatory infiltrate or to prevent such cells from migrating to the tumour site. They also might be able to 're-align' a tumour-promoting microenvironment to become a tumour-inhibiting microenvironment, to encourage tumour-specific adaptive immune responses and to inhibit metastatic spread. This potential for reversing tumour-supporting inflammation could be the start of an exciting new era for anticancer therapies. ■

1. Balkwill, F. & Mantovani, A. Inflammation and cancer: back to Virchow? *Lancet* **357,** 539–545 (2001).
This paper revisits Virchow's legacy and highlights the connections between inflammation and cancer.
2. Balkwill, F., Charles, K. A. & Mantovani, A. Smoldering and polarized inflammation in the initiation and promotion of malignant disease. *Cancer Cell* **7,** 211–217 (2005).
3. Coussens, L. M. & Werb, Z. Inflammation and cancer. *Nature* **420,** 860–867 (2002).
4. Karin, M. Nuclear factor-κB in cancer development and progression. *Nature* **441,** 431–436 (2006).
5. Koehne, C. H. & Dubois, R. N. COX-2 inhibition and colorectal cancer. *Semin. Oncol.* **31,** 12–21 (2004).
6. Flossmann, E. & Rothwell, P. M. Effect of aspirin on long-term risk of colorectal cancer: consistent evidence from randomised and observational studies. *Lancet* **369,** 1603–1613 (2007).
7. Chan, A. T., Ogino, S. & Fuchs, C. S. Aspirin and the risk of colorectal cancer in relation to the expression of COX-2. *N. Engl. J. Med.* **356,** 2131–2142 (2007).
8. Borrello, M. G. *et al.* Induction of a proinflammatory program in normal human thyrocytes by the *RET/PTC1* oncogene. *Proc. Natl Acad. Sci. USA* **102,** 14825–14830 (2005).
This is the first report that a frequent genetic event that causes cancer in humans (rearrangement of the chromosome on which *RET* is located, in human papillary thyroid carcinoma) activates an inflammatory transcriptional program in normal cells that is associated with metastatic behaviour.
9. De Falco, V. *et al.* Biological role and potential therapeutic targeting of the chemokine receptor CXCR4 in undifferentiated thyroid cancer. *Cancer Res.* **67,** 11821–11829 (2007).
10. Xu, K. & Shu, H. K. EGFR activation results in enhanced cyclooxygenase-2 expression through p38 mitogen-activated protein kinase-dependent activation of the Sp1/Sp3 transcription factors in human gliomas. *Cancer Res.* **67,** 6121–6129 (2007).
11. Guerra, C. *et al.* Chronic pancreatitis is essential for induction of pancreatic ductal adenocarcinoma by K-Ras oncogenes in adult mice. *Cancer Cell* **11,** 291–302 (2007).
12. Sparmann, A. & Bar-Sagi, D. Ras-induced interleukin-8 expression plays a critical role in tumor growth and angiogenesis. *Cancer Cell* **6,** 447–458 (2004).
13. Sumimoto, H., Imabayashi, F., Iwata, T. & Kawakami, Y. The BRAF–MAPK signaling pathway is essential for cancer-immune evasion in human melanoma cells. *J. Exp. Med.* **203,** 1651–1656 (2006).
14. Shchors, K. *et al.* The Myc-dependent angiogenic switch in tumors is mediated by interleukin 1β. *Genes Dev.* **20,** 2527–2538 (2006).
15. Soucek, L. *et al.* Mast cells are required for angiogenesis and macroscopic expansion of Myc-induced pancreatic islet tumors. *Nature Med.* **13,** 1211–1218 (2007).
16. Balkwill, F. Cancer and the chemokine network. *Nature Rev. Cancer* **4,** 540–550 (2004).
17. Kobielak, A. & Fuchs, E. Links between α-catenin, NF-κB, and squamous cell carcinoma in skin. *Proc. Natl Acad. Sci. USA* **103,** 2322–2327 (2006).
18. Phillips, R. J. *et al.* Epidermal growth factor and hypoxia-induced expression of CXC chemokine receptor 4 on non-small cell lung cancer cells is regulated by the phosphatidylinositol 3-kinase/PTEN/AKT/mammalian target of rapamycin signaling pathway and activation of hypoxia inducible factor-1α. *J. Biol. Chem.* **280,** 22473–22481 (2005).
19. Schioppa, T. *et al.* Regulation of the chemokine receptor CXCR4 by hypoxia. *J. Exp. Med.* **198,** 1391–1402 (2003).
20. Staller, P. *et al.* Chemokine receptor CXCR4 downregulated by von Hippel-Lindau tumour suppressor pVHL. *Nature* **425,** 307–311 (2003).
21. Bierie, B. & Moses, H. L. TGF-β and cancer. *Cytokine Growth Factor Rev.* **17,** 29–40 (2006).
22. Yu, H., Kortylewski, M. & Pardoll, D. Crosstalk between cancer and immune cells: role of STAT3 in the tumour microenvironment. *Nature Rev. Immunol.* **7,** 41–51 (2007).
23. Voronov, E. *et al.* IL-1 is required for tumor invasiveness and angiogenesis. *Proc. Natl Acad. Sci. USA* **100,** 2645–2650 (2003).
24. Grivennikov, S. & Karin, M. Autocrine IL-6 signaling: a key event in tumorigenesis? *Cancer Cell* **13,** 7–9 (2008).
25. Szlosarek, P. W. & Balkwill, F. R. Tumour necrosis factor α: a potential target for the therapy of solid tumours. *Lancet Oncol.* **4,** 565–573 (2003).
26. Langowski, J. L. *et al.* IL-23 promotes tumour incidence and growth. *Nature* **442,** 461–465 (2006).
27. Courtois, G. & Gilmore, T. D. Mutations in the NF-κB signaling pathway: implications for human disease. *Oncogene* **25,** 6831–6843 (2006).
28. Carbia-Nagashima, A. *et al.* RSUME, a small RWD-containing protein, enhances SUMO conjugation and stabilizes HIF-1α during hypoxia. *Cell* **131,** 309–323 (2007).
29. Mizukami, Y. *et al.* Induction of interleukin-8 preserves the angiogenic response in HIF-1α-deficient colon cancer cells. *Nature Med.* **11,** 992–997 (2005).
30. Rius, J. *et al.* NF-κB links innate immunity to the hypoxic response through transcriptional regulation of HIF-1α. *Nature* **453,** 807–811 (2008).
31. Greten, F. R. *et al.* IKKβ links inflammation and tumorigenesis in a mouse model of colitis-associated cancer. *Cell* **118,** 285–296 (2004).
32. Pikarsky, E. *et al.* NF-κB functions as a tumour promoter in inflammation-associated cancer. *Nature* **431,** 461–466 (2004).
References 31 and 32 provide evidence that NF-κB is an endogenous promoter of colon and liver carcinogenesis. Reference 31 also shows that NF-κB activation in myeloid cells is required for colitis-associated cancer.
33. Maeda, S., Kamata, H., Luo, J. L., Leffert, H. & Karin, M. IKKβ couples hepatocyte death to cytokine-driven compensatory proliferation that promotes chemical hepatocarcinogenesis. *Cell* **121,** 977–990 (2005).
34. Garlanda, C. *et al.* Increased susceptibility to colitis-associated cancer of mice lacking TIR8, an inhibitory member of the interleukin-1 receptor family. *Cancer Res.* **67,** 6017–6021 (2007).
35. Xiao, H. *et al.* The Toll–interleukin-1 receptor member SIGIRR regulates colonic epithelial homeostasis, inflammation, and tumorigenesis. *Immunity* **26,** 461–475 (2007).
36. Biswas, S. K. *et al.* A distinct and unique transcriptional program expressed by tumor-associated macrophages: defective NF-κB and enhanced IRF-3/STAT1 activation. *Blood* **107,** 2112–2122 (2006).
37. Saccani, A. *et al.* p50 nuclear factor-κB overexpression in tumor-associated macrophages inhibits M1 inflammatory responses and antitumor resistance. *Cancer Res.* **66,** 11432–11440 (2006).
38. Bromberg, J. F. *et al.* Stat3 as an oncogene. *Cell* **98,** 295–303 (1999).
39. Wang, T. *et al.* Regulation of the innate and adaptive immune responses by Stat-3 signaling in tumor cells. *Nature Med.* **10,** 48–54 (2004).
40. Kortylewski, M. *et al.* Inhibiting Stat3 signaling in the hematopoietic system elicits multicomponent antitumor immunity. *Nature Med.* **11,** 1314–1321 (2005).
41. Mantovani, A., Bottazzi, B., Colotta, F., Sozzani, S. & Ruco, L. The origin and function of tumor-associated macrophages. *Immunol. Today* **13,** 265–270 (1992).
42. Coussens, L. M., Tinkle, C. L., Hanahan, D. & Werb, Z. MMP-9 supplied by bone marrow-derived cells contributes to skin carcinogenesis. *Cell* **103,** 481–490 (2000).
43. Bunt, S. K. *et al.* Reduced inflammation in the tumor microenvironment delays the accumulation of myeloid-derived suppressor cells and limits tumor progression. *Cancer Res.* **67,** 10019–10026 (2007).
44. Lin, E. Y., Nguyen, A. V., Russell, R. G. & Pollard, J. W. Colony-stimulating factor 1 promotes progression of mammary tumors to malignancy. *J. Exp. Med.* **193,** 727–740 (2001).
This paper describes the first genetic evidence that TAMs promote cancer, in a study of a primary breast carcinoma model.
45. De Palma, M. *et al.* Tie2 identifies a hematopoietic lineage of proangiogenic monocytes required for tumor vessel formation and a mesenchymal population of pericyte progenitors. *Cancer Cell* **8,** 211–226 (2005).
46. Sica, A. & Bronte, V. Altered macrophage differentiation and immune dysfunction in tumor development. *J. Clin. Invest.* **117,** 1155–1166 (2007).
47. Mantovani, A., Sozzani, S., Locati, M., Allavena, P. & Sica, A. Macrophage polarization: tumor-associated macrophages as a paradigm for polarized M2 mononuclear phagocytes. *Trends Immunol.* **23,** 549–555 (2002).
48. Hagemann, T. *et al.* Ovarian cancer cells polarize macrophages toward a tumor-associated phenotype. *J. Immunol.* **176,** 5023–5032 (2006).
49. Fischer, C. *et al.* Anti-PlGF inhibits growth of VEGFR-inhibitor-resistant tumors without affecting healthy vessels. *Cell* **131,** 463–475 (2007).
50. Kaplan, R. N. *et al.* VEGFR1-positive haematopoietic bone marrow progenitors initiate the pre-metastatic niche. *Nature* **438,** 820–827 (2005).
51. Shojaei, F. *et al.* Bv8 regulates myeloid-cell-dependent tumour angiogenesis. *Nature* **450,** 825–831 (2007).
52. Coussens, L. M. *et al.* Inflammatory mast cells up-regulate angiogenesis during squamous epithelial carcinogenesis. *Genes Dev.* **13,** 1382–1397 (1999).
53. Lewis, C. E., De Palma, M. & Naldini, L. Tie2-expressing monocytes and tumor angiogenesis: regulation by hypoxia and angiopoietin 2. *Cancer Res.* **67,** 8429–8432 (2007).
54. Sozzani, S., Rusnati, M., Riboldi, E., Mitola, S. & Presta, M. Dendritic cell-endothelial cell cross-talk in angiogenesis. *Trends Immunol.* **28,** 385–392 (2007).
55. Noonan, D. M., De Lerma Barbaro, A., Vannini, N., Mortara, L. & Albini, A. Inflammation, inflammatory cells and angiogenesis: decisions and indecisions. *Cancer Metastasis Rev.* **27,** 31–40 (2008).
56. Dunn, G. P., Old, L. J. & Schreiber, R. D. The immunobiology of cancer immunosurveillance and immunoediting. *Immunity* **21,** 137–148 (2004).
57. Swann, J. B. *et al.* Demonstration of inflammation-induced cancer and cancer immunoediting during primary tumorigenesis. *Proc. Natl Acad. Sci. USA* **105,** 652–656 (2008).
58. de Visser, K. E., Korets, L. V. & Coussens, L. M. De novo carcinogenesis promoted by chronic inflammation is B lymphocyte dependent. *Cancer Cell* **7,** 411–423 (2005).
This paper shows that in a model of human-papilloma-virus-driven carcinogenesis, adaptive immune responses mediated by B cells coordinate cancer-promoting inflammation.
59. Beatson, G. On the treatment of inoperable cases of carcinoma of the mamma: suggestions for a new method of treatment with illustrative cases. *Lancet* **2,** 104–162, (1896).
60. Zhu, P. *et al.* Macrophage/cancer cell interactions mediate hormone resistance by a nuclear receptor derepression pathway. *Cell* **124,** 615–629 (2006).
61. Naugler, W. E. *et al.* Gender disparity in liver cancer due to sex differences in MyD88-dependent IL-6 production. *Science* **317,** 121–124 (2007).
References 60 and 61 show that two classic pathways of cancer promotion, hormones and inflammation, are linked in both liver cancer and prostate cancer.
62. Muller, A. *et al.* Involvement of chemokine receptors in breast cancer metastasis. *Nature* **410,** 50–56 (2001).
63. Burger, J. A. & Kipps, T. J. CXCR4: a key receptor in the crosstalk between tumor cells and their microenvironment. *Blood* **107,** 1761–1767 (2006).
64. Kaifi, J. T. *et al.* Tumor-cell homing to lymph nodes and bone marrow and CXCR4 expression in esophageal cancer. *J. Natl Cancer Inst.* **97,** 1840–1847 (2005).
65. Salvucci, O. *et al.* The role of CXCR4 receptor expression in breast cancer: a large tissue microarray study. *Breast Cancer Res. Treat.* **97,** 275–283 (2006).
66. Kim, J. *et al.* Chemokine receptor CXCR4 expression in colorectal cancer patients increases the risk for recurrence and for poor survival. *J. Clin. Oncol.* **23,** 2744–2753 (2005).

67. Shields, J. D. *et al.* Autologous chemotaxis as a mechanism of tumor cell homing to lymphatics via interstitial flow and autocrine CCR7 signaling. *Cancer Cell* **11,** 526–538 (2007).
68. Kawada, K. *et al.* Pivotal role of CXCR3 in melanoma cell metastasis to lymph nodes. *Cancer Res.* **64,** 4010–4017 (2004).
69. Shulby, S. A., Dolloff, N. G., Stearns, M. E., Meucci, O. & Fatatis, A. CX_3CR1-fractalkine expression regulates cellular mechanisms involved in adhesion, migration, and survival of human prostate cancer cells. *Cancer Res.* **64,** 4693–4698 (2004).
70. Burns, J. M. *et al.* A novel chemokine receptor for SDF-1 and I-TAC involved in cell survival, cell adhesion, and tumor development. *J. Exp. Med.* **203,** 2201–2213 (2006).
71. Zipin-Roitman, A. *et al.* CXCL10 promotes invasion-related properties in human colorectal carcinoma cells. *Cancer Res.* **67,** 3396–3405 (2007).
72. Ghadjar, P. *et al.* Chemokine receptor CCR6 expression level and liver metastases in colorectal cancer. *J. Clin. Oncol.* **24,** 1910–1916 (2006).
73. Kulbe, H. *et al.* The inflammatory cytokine tumor necrosis factor-α generates an autocrine tumor-promoting network in epithelial ovarian cancer cells. *Cancer Res.* **67,** 585–592 (2007).
74. Kulbe, H., Hagemann, T., Szlosarek, P. W., Balkwill, F. R. & Wilson, J. L. The inflammatory cytokine tumor necrosis factor-α regulates chemokine receptor expression on ovarian cancer cells. *Cancer Res.* **65,** 10355–10362 (2005).
75. Bates, R. C. & Mercurio, A. M. Tumor necrosis factor-α stimulates the epithelial-to-mesenchymal transition of human colonic organoids. *Mol. Biol. Cell* **14,** 1790–1800 (2003).
76. Luo, J. L. *et al.* Nuclear cytokine-activated IKKα controls prostate cancer metastasis by repressing maspin. *Nature* **446,** 690–694 (2007).
77. Condeelis, J. & Pollard, J. W. Macrophages: obligate partners for tumor cell migration, invasion, and metastasis. *Cell* **124,** 263–266 (2006).
78. Wyckoff, J. B. *et al.* Direct visualization of macrophage-assisted tumor cell intravasation in mammary tumors. *Cancer Res.* **67,** 2649–2656 (2007).
79. Robinson-Smith, T. M. *et al.* Macrophages mediate inflammation-enhanced metastasis of ovarian tumors in mice. *Cancer Res.* **67,** 5708–5716 (2007).
80. Hagemann, T. *et al.* Macrophages induce invasiveness of epithelial cancer cells via NF-κB and JNK. *J. Immunol.* **175,** 1197–1205 (2005).
81. Marchesi, F. *et al.* Increased survival, proliferation, and migration in metastatic human pancreatic tumor cells expressing functional CXCR4. *Cancer Res.* **64,** 8420–8427 (2004).
82. Nickoloff, B. J., Ben-Neriah, Y. & Pikarsky, E. Inflammation and cancer: is the link as simple as we think? *J. Invest. Dermatol.* **124,** x-xiv (2005).
83. Hagemann, T. *et al.* Re-educating tumor-associated macrophages by targeting NF-κB. *J. Exp. Med.* **205,** 1261–1268 (2008).
This paper shows that NF-κB activated through the IL-1 receptor and MyD88 signalling pathway maintains the phenotype of TAMs, suggesting that tumour-promoting macrophages might be re-educated by the targeting of NF-κB.
84. Dajee, M. *et al.* NF-κB blockade and oncogenic Ras trigger invasive human epidermal neoplasia. *Nature* **421,** 639–643 (2003).
85. Coley, W. B. The treatment of malignant tumors by repeated inoculations of erysipelas: with a report of ten original cases. *Am. J. Med. Sci.* **105,** 487–511 (1893).
86. Apetoh, L. *et al.* Toll-like receptor 4-dependent contribution of the immune system to anticancer chemotherapy and radiotherapy. *Nature Med.* **13,** 1050–1059 (2007).
87. Sato, Y. *et al.* IL-10 deficiency leads to somatic mutations in a model of IBD. *Carcinogenesis* **27,** 1068–1073 (2006).
88. Bielas, J. H., Loeb, K. R., Rubin, B. P., True, L. D. & Loeb, L. A. Human cancers express a mutator phenotype. *Proc. Natl Acad. Sci. USA* **103,** 18238–18242 (2006).
89. Gungor, N., Godschalk, R. W. L., Pachen, D. M., Van Schooten, F. J. & Knaapen, A. M. Activated neutrophils inhibit nucleotide excision repair in human pulmonary epithelial cells: role of myeloperoxidase. *FASEB J.* **21,** 2359–2367 (2007).
90. Dally, H. *et al.* Myeloperoxidase (MPO) genotype and lung cancer histologic types: the MPO –463 A allele is associated with reduced risk for small cell lung cancer in smokers. *Int. J. Cancer* **102,** 530–535 (2002).
91. Rao, V. P. *et al.* Innate immune inflammatory response against enteric bacteria *Helicobacter hepaticus* induces mammary adenocarcinoma in mice. *Cancer Res.* **66,** 7395–7400 (2006).
92. Mantovani, A. Cancer: an infernal triangle. *Nature* **448,** 547–548 (2007).
93. Madhusudan, S. *et al.* Study of etanercept, a tumor necrosis factor-α inhibitor, in recurrent ovarian cancer. *J. Clin. Oncol.* **23,** 5950–5959 (2005).
94. Brown, E. R. *et al.* A clinical study assessing the tolerability and biological effects of infliximab, a TNF-α inhibitor, in patients with advanced cancer. *Ann. Oncol.* **19,** 1340–1346 (2008).
95. Harrison, M. L. *et al.* Tumor necrosis factor α as a new target for renal cell carcinoma: two sequential phase II trials of infliximab at standard and high dose. *J. Clin. Oncol.* **25,** 4542–4549 (2007).
The paper reports the first clinical evidence that TNF-α could be a target for treating renal-cell carcinoma.
96. Weber, D. M. *et al.* Lenalidomide plus dexamethasone for relapsed multiple myeloma in North America. *N. Engl. J. Med.* **357,** 2133–2142 (2007).
97. Bertagnolli, M. M. *et al.* Celecoxib for the prevention of sporadic colorectal adenomas. *N. Engl. J. Med.* **355,** 873–884 (2006).
98. Steinbach, G. *et al.* The effect of celecoxib, a cyclooxygenase-2 inhibitor, in familial adenomatous polyposis. *N. Engl. J. Med.* **342,** 1946–1952 (2000).

Acknowledgements A.M., P.A. and A.S. are supported by the Italian Association for Cancer Research, the Italian Ministry of Health, the Italian Ministry of Universities and Research, and the European Commission. F.B. is supported by Cancer Research UK, the Medical Research Council, the Association for International Cancer Research and the Higher Education Funding Council for England.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence should be addressed to A.M. (alberto.mantovani@humanitas.it).

The development of allergic inflammation

Stephen J. Galli[1,2], Mindy Tsai[1] & Adrian M. Piliponsky[1]

Allergic disorders, such as anaphylaxis, hay fever, eczema and asthma, now afflict roughly 25% of people in the developed world. In allergic subjects, persistent or repetitive exposure to allergens, which typically are intrinsically innocuous substances common in the environment, results in chronic allergic inflammation. This in turn produces long-term changes in the structure of the affected organs and substantial abnormalities in their function. It is therefore important to understand the characteristics and consequences of acute and chronic allergic inflammation, and in particular to explore how mast cells can contribute to several features of this maladaptive pattern of immunological reactivity.

The conception that antibodies, which should protect against disease, are also responsible for disease, sounds at first absurd.
Clemens von Pirquet (1906)

The term 'allergy' was coined by Clemens von Pirquet in 1906 to call attention to the unusual propensity of some individuals to develop signs and symptoms of reactivity, or 'hypersensitivity reactions', when exposed to certain substances[1] (Box 1). Although the statement quoted above pertained to the cause of serum sickness[2], allergic disorders (also known as atopic disorders, from the Greek *atopos*, meaning out of place) are also associated with the production of allergen-specific IgE and with the expansion of allergen-specific T-cell populations, both of which are reactive with what typically are otherwise harmless environmental substances. These disorders are increasingly prevalent in the developed world and include allergic rhinitis (also known as hay fever), atopic dermatitis (also known as eczema), allergic (or atopic) asthma and some food allergies[3–5]. Some people develop a potentially fatal systemic allergic reaction, termed anaphylaxis, within seconds or minutes of exposure to allergens[6].

In recent years, it has become clear that much of the pathology, and therefore the burden of disease, associated with allergic disorders reflects the long-term consequences of chronic allergic inflammation at sites of persistent or repetitive exposure to allergens[3,4] (Box 1). This realization has led to renewed efforts to define additional therapeutic targets in allergic disease[7–9], to devise improved strategies to induce immunological tolerance to the offending allergens[10,11], and even to manipulate the immune response to prevent the initial development of allergic disorders[12].

Here we outline some of the factors that can contribute to the development of IgE-associated allergic disorders and describe the features of allergic inflammation. We focus on the effects of short-term and long-term allergic inflammation on the structure and function of the affected tissues, particularly in asthma, and on the evidence that mast cells can contribute to multiple features of chronic, as well as acute, allergic inflammation. Finally, we briefly consider some of the approaches that are being used or contemplated to manage disorders associated with allergic inflammation. Some other disorders can also be considered allergic, such as allergic contact dermatitis and hypersensitivity pneumonitis, but these do not develop by the same immunological mechanisms — that is, they do not involve IgE- and T helper 2 (T_H2)-cell mediated responses[4] — and therefore are not discussed here.

Allergy and gene–environment interactions

Many features of allergic inflammation resemble those of the inflammation that results from immune responses to infection with enteric helminths[13] or from cutaneous responses to the bites of ectoparasites such as ticks[14]. Similarities to aspects of immune responses to parasites or environmental allergens have also been identified, notably that both involve T_H2 cells and are associated with antigen-specific IgE. These similarities have led to the idea that in allergic disorders the immune system is 'tricked' into reacting to otherwise inconsequential allergens in the same way as it does to signals derived from enteric helminths or ectoparasites.

In addition to the benefits conferred on the host by T_H2-cell responses to parasites, such as the development and enhancement of effector mechanisms that contribute to parasite clearance, chronic infection with certain parasites often also turns on immunological mechanisms that downregulate the inflammation and tissue damage that is associated with that infection[13,15]. Such mechanisms include the development of regulatory T cells that secrete interleukin 10 (IL-10), which has many immunosuppressive and anti-inflammatory effects[13,15,16]. In allergic disorders, it is thought that such downregulatory mechanisms do not fully develop, are lost or might be overwhelmed by inflammatory factors[13,15,16]. Indeed, observations of this type support the 'hygiene hypothesis'[5,13,15,16]. This hypothesis is based on the observation that, as living standards advance, there is reduced exposure to parasitic infections and to other pathogenic and non-pathogenic microorganisms (and their products). Such infections usually promote the normal development of immune responses (with a bias towards T_H1 cells rather than T_H2 cells) and favour the development of appropriate control of potentially harmful immune responses by various populations of regulatory T cells. However, as exposure to infections is reduced, and exposure to certain otherwise harmless environmental allergens is increased, there is a propensity for genetically predisposed individuals to develop T_H2-cell-type responses to a variety of common environmental allergens[5,13,15,16].

The molecular mechanisms underlying the hygiene hypothesis continue to be explored[13,15–17], but there can be no doubt that the recent marked increase in allergic disorders reflects recent changes in the interactions between the external environment and those individuals who are genetically predisposed to develop allergic diseases. Accordingly, many researchers are attempting to understand the gene–environment interactions that promote the development, increase the severity or limit the resolution of allergic inflammation[18,19]. There is already evidence that

[1]Department of Pathology, Stanford University School of Medicine, 300 Pasteur Drive, Stanford, California 94305, USA. [2]Department of Microbiology and Immunology, Stanford University School of Medicine, 300 Pasteur Drive, Stanford, California 94305, USA.

Box 1 | Defining allergy, allergens and allergic inflammation

The term allergy can be used to refer to abnormal adaptive immune responses that either involve or do not involve allergen-specific IgE. This Review focuses on the former: that is, on the development, characteristics and consequences of the allergic inflammation that occurs in disorders in which IgE is thought to participate.

Allergy
An abnormal adaptive immune response directed against non-infectious environmental substances (allergens), including non-infectious components of certain infectious organisms. In allergic disorders, such as anaphylaxis, allergic rhinitis (hay fever), some food allergies and allergic asthma, these responses are characterized by the involvement of allergen-specific IgE and T helper 2 (T_H2) cells that recognize allergen-derived antigens. In other kinds of allergy, such as allergic contact dermatitis, IgE is thought not to be important.

Allergen
There are two main types of allergen.

The first type encompasses any non-infectious environmental substance that can induce IgE production (thereby 'sensitizing' the subject) so that later re-exposure to that substance induces an allergic reaction. Common sources of allergens include grass and tree pollens, animal dander (sheddings from skin and fur), house-dust-mite faecal particles, certain foods (notably peanuts, tree nuts, fish, shellfish, milk and eggs), latex, some medicines and insect venoms. In some instances, allergen-specific IgE directed against foreign antigens can also recognize crossreactive host antigens, but the clinical significance of this is unclear.

The second type is a non-infectious environmental substance that can induce an adaptive immune response associated with local inflammation but is thought to occur independently of IgE (for example, allergic contact dermatitis to poison ivy or nickel).

Allergic inflammation
The inflammation produced in sensitized subjects after exposure to a specific allergen(s). A single allergen exposure produces an acute reaction, which is known as an early-phase reaction or a type I immediate hypersensitivity reaction. In many subjects, this is followed by a late-phase reaction. With persistent or repetitive exposure to allergen, chronic allergic inflammation develops, with associated tissue alterations.

Early-phase reaction
An IgE-mediated type I immediate hypersensitivity reaction that can occur within minutes of allergen exposure. Reactions can be localized (for example, acute rhinoconjunctivitis in allergic rhinitis, acute asthma attacks, urticaria (hives) and gastrointestinal reactions in food allergies) or systemic (anaphylaxis). In such reactions, IgE bound to FcεRI on mast cells and basophils is crosslinked by allergen, resulting in the release of the cells' diverse preformed and newly synthesized mediators. These events cause vasodilation, increased vascular permeability with oedema, and acute functional changes in affected organs (such as bronchoconstriction, airway mucus secretion, urticaria, vomiting and diarrhoea). Some of the released mediators also promote the local recruitment and activation of leukocytes, contributing to the development of late-phase reactions.

Late-phase reaction
A reaction that typically develops after 2–6 h and peaks 6–9 h after allergen exposure. It is usually preceded by a clinically evident early-phase reaction and fully resolves in 1–2 days. Skin late-phase reactions involve oedema, pain, warmth and erythema (redness). In the lungs, these reactions are characterized by airway narrowing and mucus hypersecretion. They reflect the local recruitment and activation of T_H2 cells, eosinophils, basophils and other leukocytes, and persistent mediator production by resident cells (such as mast cells). Mediators that initiate late-phase reactions are thought to be derived from resident mast cells activated by IgE and allergen or from T cells that recognize allergen-derived peptides (such T cells may be either resident at, or recruited to, sites of allergen challenge).

Chronic allergic inflammation
Persistent inflammation induced by prolonged or repetitive exposure to specific allergens, typically characterized not only by the presence of large numbers of innate and adaptive immune cells (in the form of leukocytes) at the affected site but also by substantial changes in the extracellular matrix and alterations in the number, phenotype and function of structural cells in the affected tissues.

exposure to the same microbial products can have the opposite effect on an individual's propensity to develop allergic disorders, depending on an individual's genotype[19].

Allergen sensitization and epithelial barriers

Sensitization to an allergen reflects the allergen's ability to elicit a T_H2-cell response, in which IL-4 and IL-13 drive IgE production by promoting immunoglobulin class-switch recombination in B cells[4,10,11,20,21] (Fig. 1).

Many factors affect the likelihood of developing clinically significant sensitization[18,19]: host genotype, type of allergen, allergen concentration in the environment and whether exposure occurs together with agents that can enhance the sensitization process. These agents include certain ligands of Toll-like receptors, including endotoxin, which can promote T_H1-cell responses (as proposed in the hygiene hypothesis) and in certain circumstances (such as when encountered in appropriate concentrations together with an allergen) might be able to enhance the development of T_H2-cell responses[22]. Other agents that can enhance allergic sensitization are chitin, which is found in many organisms (including some that are important sources of allergens[23]), and environmental pollutants[24]. Another important factor is the pattern of contact of the immune system with allergens: for example, the amount, frequency and/or route of allergen exposure; and the type (myeloid and/or plasmacytoid) and phenotypic characteristics of the dendritic-cell subpopulations that participate in the responses[25]. The pattern of contact may affect whether there is a strong T_H2-cell response (and therefore clinical allergy), a T_H2-cell response that is kept in check by IL-10-secreting, and perhaps other, regulatory T cells[10,11,16], a modified T_H2-cell response that results in high concentrations of allergen-specific IgG_4 (ref. 26) or another form of immunological tolerance[25].

Genetic or environmental factors that influence the epithelium, including its permeability to allergens, can favour the subsequent development of a T_H2-cell response[18,27,28]. For example, loss-of-function mutations in *FLG*, which encodes filaggrin (a protein that promotes the organization of intermediate filaments of squamous cells into bundles for later crosslinking by transglutaminases), diminish the barrier function of the skin and result in ichthyosis vulgaris, a skin disease that is inherited in a semidominant pattern with incomplete penetrance[27]. Many patients with ichthyosis vulgaris also develop atopic dermatitis, and inheriting a single copy of certain loss-of-function alleles of *FLG* is associated with a markedly increased risk of developing atopic dermatitis[27]. Such *FLG* mutations have been identified in approximately 10% of subjects of European ancestry and may occur in as many as 50% of patients who develop atopic dermatitis[27].

Patients with *FLG* mutations and atopic dermatitis are at greatly increased risk of developing asthma, even though filaggrin protein expression has so far not been detected in the lungs[29]. This finding strongly suggests that a defect in epithelial barrier function that increases the likelihood of sensitization to allergens encountered in the skin and upper airway, for example, can contribute to the development of systemic immune responses that result in allergic disease at other sites exposed to that allergen, such as the lungs. Genome-wide screens of individuals with atopic dermatitis and/or asthma have identified many other candidate genes that are expressed in the relevant epithelial-cell populations at the affected site, suggesting that mutations or polymorphisms that alter the normal barrier (and other) functions of epithelia may contribute to the development of allergies and allergic inflammation[18,27].

Most allergens are proteins (some are lipids or carbohydrates), and many, including the major house-dust-mite allergen, Der p 1, are proteases[25].

Some of these proteases can directly reduce epithelial barrier function[30] or hydrolyse substrates that participate in the development of T_H2-cell responses, including CD23, CD25, CD40 and DC-SIGN (dendritic-cell-specific ICAM3-grabbing non-integrin)[25]. Proteases are also used by parasites to invade tissues[31], and recent work suggests that basophils activated by exogenous proteases are one potential source of both thymic stromal lymphopoietin (TSLP), which can promote allergic inflammation, and the 'early IL-4' that can initiate sensitization for the development of T_H2-cell- and IgE-mediated immune responses to allergens or parasites[32,33].

Features of allergic inflammation

Allergic inflammation often is classified into three temporal phases. Early-phase reactions are induced within seconds to minutes of allergen challenge, and late-phase reactions occur within several hours. By contrast, chronic allergic inflammation is a persistent inflammation that occurs at sites of repeated allergen exposure[4,9] (Box 1).

Early-phase reactions

Early-phase reactions (or type I immediate hypersensitivity reactions[4]) occur within minutes of allergen exposure and mainly reflect the secretion of mediators by mast cells at the affected site. In sensitized individuals, these mast cells already have allergen-specific IgE bound to their surface high-affinity IgE receptors (FcεRI). When crosslinking of adjacent IgE molecules by bivalent or multivalent allergen occurs, aggregation of FcεRI triggers a complex intracellular signalling process that results in the secretion of three classes of biologically active product: those stored in the cytoplasmic granules, lipid-derived mediators, and newly synthesized cytokines, chemokines and growth factors, as well as other products[8,34–37] (Fig. 2).

The secretion of preformed mediators occurs when the membrane of the mast cells' cytoplasmic granules fuses with the plasma membrane in a process called degranulation (or compound exocytosis)[38], exposing the granule contents to the external environment. The released mediators include biogenic amines (histamine and little or no serotonin in humans, but both histamine and serotonin in mice and rats[35,36]), serglycin proteoglycans (such as heparin and chondroitin sulphate), serine proteases (such as tryptases, chymases and carboxypeptidases)[39–41], and various other enzymes and certain cytokines and growth factors that can be associated with the granules (such as tumour-necrosis factor-α (TNF-α) and vascular endothelial growth factor A (VEGFA))[35,36,42,43]. Mast cells activated by the aggregation of FcεRI also release lipid-derived mediators. They metabolize arachidonic acid through the cyclooxygenase and lipoxygenase pathways, resulting in the release of prostaglandins (particularly prostaglandin D_2 (PGD_2)), leukotriene B_4 (LTB_4) and cysteinyl leukotrienes (cys-LTs, particularly LTC_4)[44]. Some activated mast cells can also release platelet-activating factor (PAF)[45]. Both the phenotypic characteristics of mast cells (such as their mediator content and their susceptibility to activation by various stimuli) can vary considerably between mast-cell populations at different anatomical sites or as a result of exposure to cytokines or other microenvironmental factors at sites of immune responses[34,35,38–44].

The release of preformed and lipid-derived mediators contributes to the acute signs and symptoms associated with early-phase reactions[4,35,46]

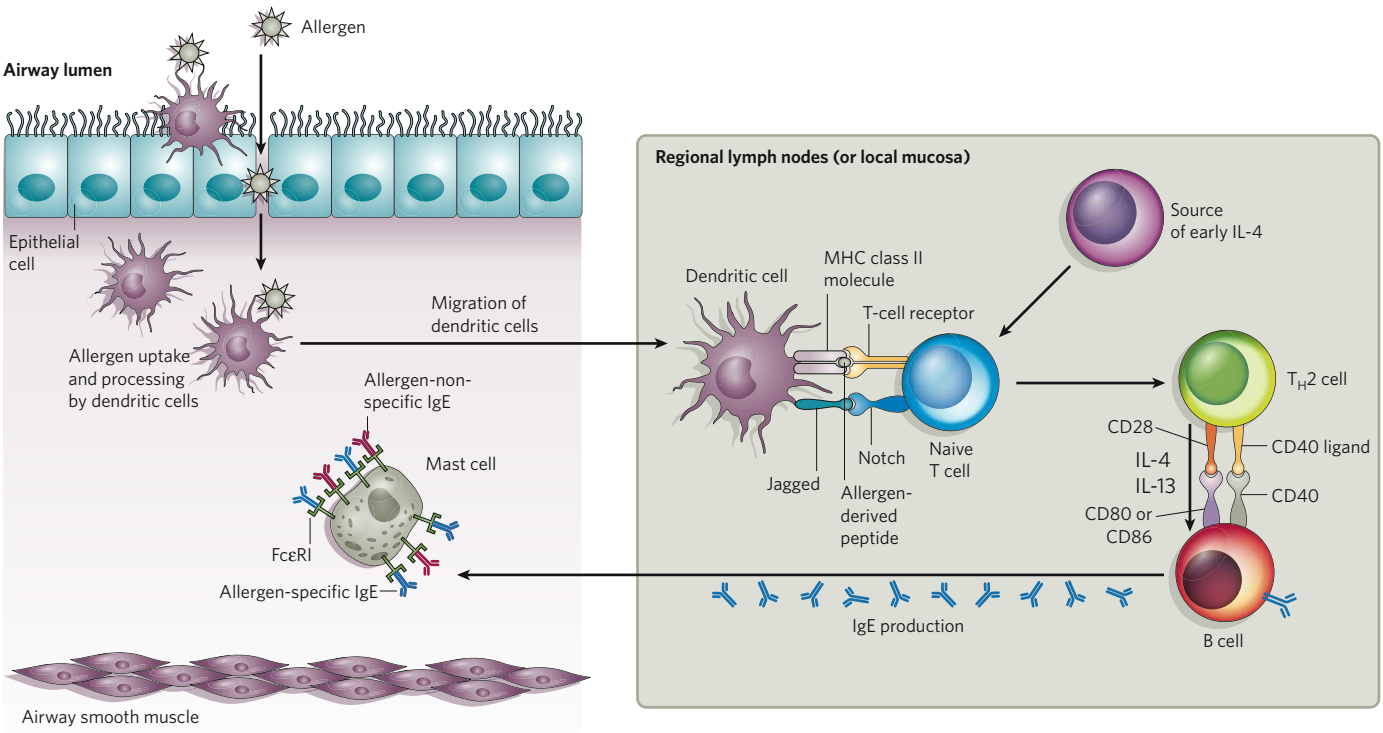


Figure 1 | Sensitization to allergens in the airway. Allergen can be sampled by dendritic cells in the airway lumen, and can enter tissues through disrupted epithelium (not shown) or, for some allergens with protease activity, can gain access to submucosal dendritic cells by cleaving epithelial-cell tight junctions. Activated dendritic cells mature and migrate to regional lymph nodes or to sites in the local mucosa, where they present peptides derived from the processed allergen in the context of major histocompatibility complex (MHC) class II molecules to naive T cells. In the presence of 'early interleukin 4' (IL-4) (potentially derived from a range of cells, including basophils, mast cells, eosinophils, natural killer T cells and T cells), naive T cells acquire the characteristics of T helper 2 (T_H2) cells, a process that may be enhanced by engagement of Notch at the surface of T cells with Jagged on dendritic cells). T_H2 cells produce IL-4 and IL-13. In the presence of these cytokines and the ligation of suitable co-stimulatory molecules (CD40 with CD40 ligand, and CD80 or CD86 with CD28), B cells undergo immunoglobulin class-switch recombination, in which the gene segments that encode the immunoglobulin heavy chain are rearranged such that antibody of the IgE class is produced. Basophils and mast cells also can produce IL-4 and/or IL-13, and can stimulate B cells through CD40 (not shown). IgE diffuses locally and enters the lymphatic vessels. It subsequently enters the blood and is then distributed systemically. After gaining access to the interstitial fluid, allergen-specific or non-specific IgE binds to the high-affinity receptor for IgE (FcεRI) on tissue-resident mast cells, thereby sensitizing them to respond when the host is later re-exposed to the allergen. Sensitization does not produce symptoms (for example, if sensitization occurs by way of the airways, bronchoconstriction does not occur). This T_H2-cell response to allergen can be downregulated or modified by various mechanisms (not shown).

(Fig. 3). These signs and symptoms vary according to the site of the reaction but can include vasodilation (in part reflecting the action of mediators on local nerves, and producing erythema (reddening) of the skin or conjunctiva), markedly increased vascular permeability (leading to tissue swelling and, in the eyes, tear formation), contraction of bronchial smooth muscle (producing airflow obstruction and wheezing), and increased secretion of mucus (exacerbating airflow obstruction in the lower airways and producing a runny nose). Such mediators can also stimulate nociceptors of sensory nerves (both C-fibre-type unmyelinated nerves and thinly myelinated Aδ nerves) of the nose[47], skin[48] and airway[49], resulting in sneezing, itching or coughing.

When such mediators are released locally, an early-phase reaction ensues. By contrast, the rapid and systemic release of such mediators, from mast cells and basophils (which also express FcεRI and can release a panel of mediators similar, but not identical, to those of mast cells[35,50,51]), accounts for much of the pathology associated with anaphylaxis[6].

Late-phase reactions

Mast cells responding to IgE and allergen also release a broad range of newly synthesized cytokines, chemokines and growth factors[8,35-37], but these are released more slowly than the preformed mediators. Some mast-cell populations also can rapidly secrete some of these products, including TNF-α, from preformed stores[35]. Some mast-cell products have the potential to recruit other immune cells either directly or indirectly (for example, TNF-α, LTB_4, IL-8 (also known as CXCL8), CC-chemokine ligand 2 (CCL2) and many other chemokines), to activate innate immune cells (for example, TNF-α and IL-5), and to affect many aspects of the biology of dendritic cells, T cells and B cells (for example, IL-10, TNF-α, transforming growth factor-β (TGF-β) and histamine)[35,52]. However, some products secreted by activated mast cells (such as IL-10 and TGF-β) can have anti-inflammatory or immunosuppressive functions[52,53]. Certain mast-cell-derived products can also influence the biology of structural cells, including vascular endothelial cells, epithelial cells, fibroblasts, smooth muscle cells and nerve cells[28,39–41,44,54,55]. Other products that contribute to late-phase reactions can be derived from T cells that recognize allergen-derived peptides; such T cells may be either resident at or recruited to early-phase reactions at sites of allergen challenge[4,9,56].

Late-phase reactions (Fig. 4) are thought to be coordinated in part by certain long-term consequences of the mediators released by activated mast cells during early-phase reactions, and in part by antigen-stimulated T cells. The clinical features of late-phase reactions reflect the activities

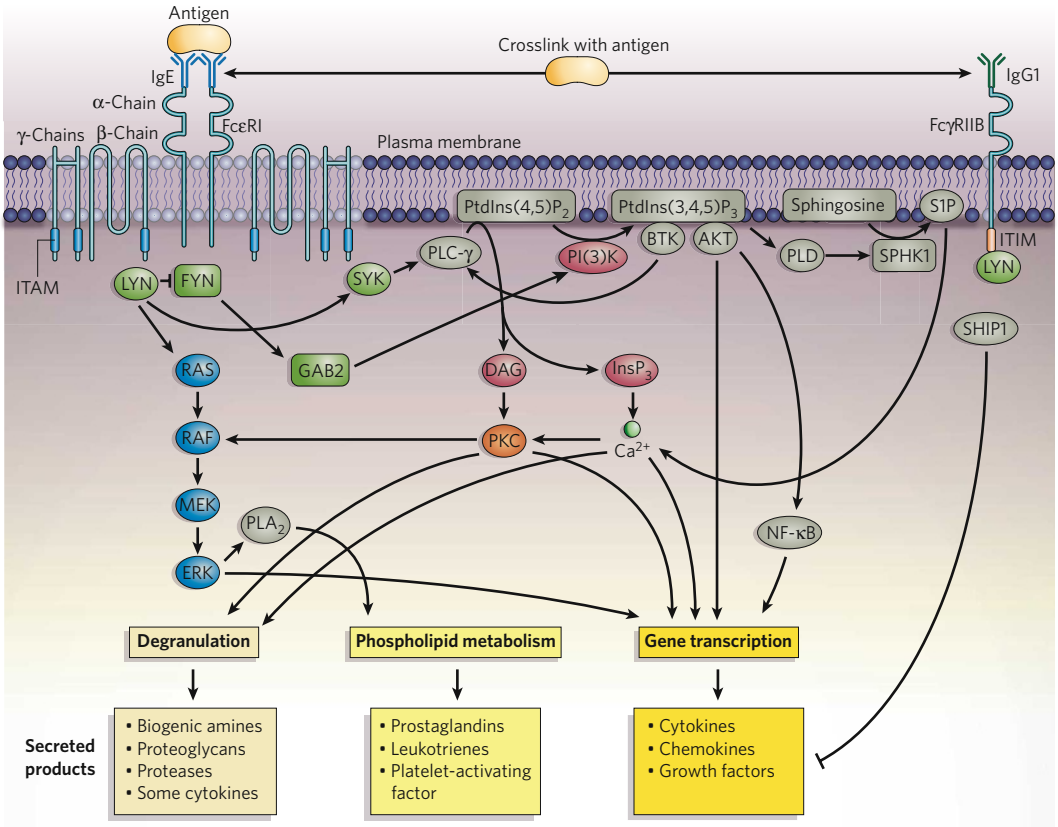


Figure 2 | Highly simplified scheme of FcεRI signalling events in mast cells. Crosslinking of FcεRI-bound IgE with antigen induces aggregation of two or more FcεRI molecules and activates the protein tyrosine kinases LYN and FYN. LYN, in turn, phosphorylates the immunoreceptor tyrosine-based activation motifs (ITAMs) in FcεRI and activates the protein tyrosine kinase SYK (after SYK has bound to an ITAM). FYN phosphorylates the adaptor GAB2, activating the phosphatidylinositol-3-OH kinase (PI(3)K) pathway. LYN and SYK phosphorylate many adaptor molecules (such as LAT, not shown) and enzymes, thereby regulating the activation of the RAS–MAPK (mitogen-activated protein kinase), phospholipase C-γ (PLC-γ) and PI(3)K pathways, as well as other pathways. (LYN also can negatively regulate FYN activity.) The RAS–MAPK pathway — a protein-kinase cascade that involves RAS, RAF, MEK and ERK — activates transcription factors (thereby regulating the synthesis of protein mediators) and activates PLA_2, which participates in arachidonic acid metabolism (thereby regulating the production of lipid-derived mediators). PLC-γ activation regulates calcium (Ca^{2+}) responses, by generating inositol-1,4,5-trisphosphate ($InsP_3$), and protein kinase C (PKC) activation, by generating diacylglycerol (DAG). The PI(3)K product phosphatidylinositol-3,4,5-trisphosphate ($PtdIns(3,4,5)P_3$) is an important lipid mediator that regulates the formation of other lipid mediators, such as DAG and sphingosine 1-phosphate (S1P), and the activity of various enzymes, such as Bruton's tyrosine kinase (BTK) and AKT. FcεRI can be induced to co-aggregate with FcγRIIB (a low-affinity receptor for IgG), for example when IgE and IgG1 are bound to the same antigen. This process inhibits FcεRI signalling events, and therefore mast-cell activation and product secretion, through the LYN-mediated phosphorylation of the FcγRIIB ITIM (immunoreceptor tyrosine-based inhibitory motif) and the recruitment of the inositol phosphatase SHIP1 (which catalyses the hydrolysis of $PtdIns(3,4,5)P_3$ to $PtdIns(3,4)P_2$) (not shown). Some arrows do not indicate direct interactions or targets. NF-κB, nuclear factor-κB; SPHK1, sphingosine kinase 1.

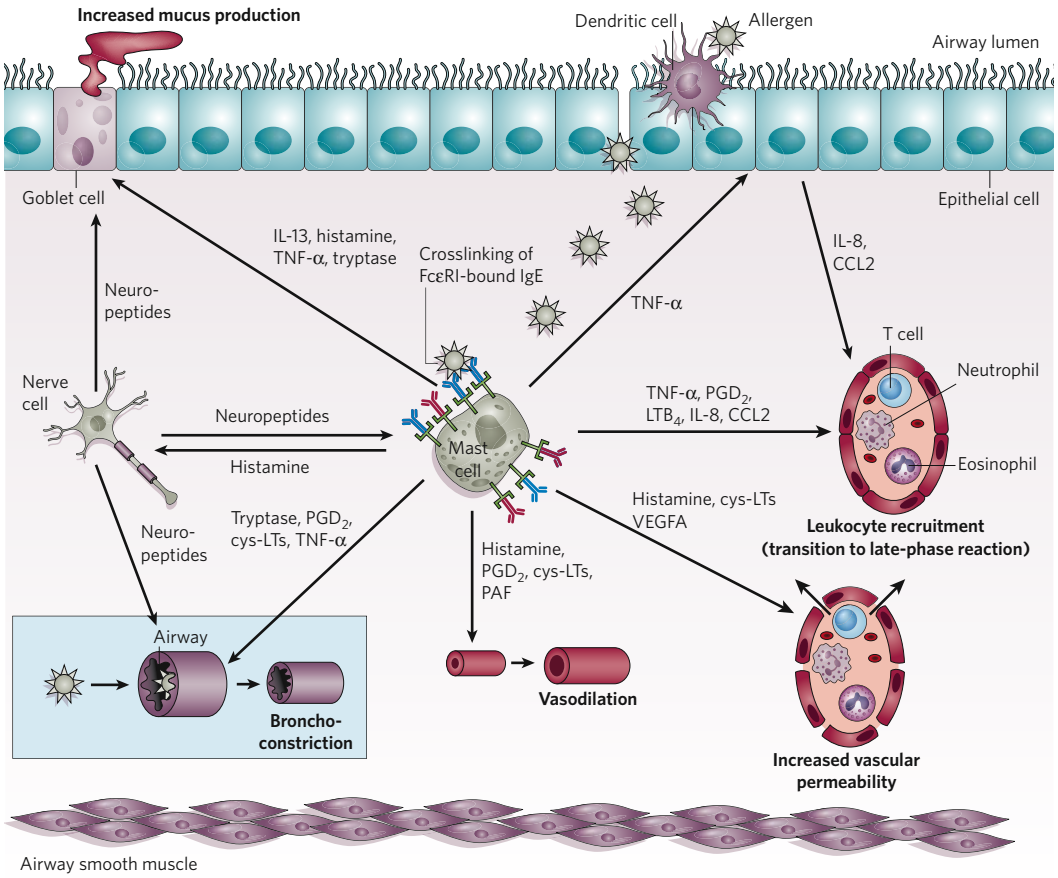


Figure 3 | Early phase of allergen-induced airway inflammation. The individual IgE molecules that are bound to the FcεRI molecules on a single mast cell can be specific for different antigens. The recognition of a particular allergen by FcεRI-bound IgE specific for antigen derived from that allergen (allergen-specific IgE) induces FcεRI aggregation, which activates mast cells to secrete preformed mediators and lipid-derived mediators and to increase the synthesis of many cytokines, chemokines and growth factors. The rapidly secreted mediators result in bronchoconstriction (lower left), vasodilation, increased vascular permeability and increased mucus production. Mast cells also contribute to the transition to the late-phase reaction (Fig. 4) by promoting an influx of inflammatory leukocytes, both by upregulating adhesion molecules on vascular endothelial cells (for example, through TNF-α) and by secreting chemotactic mediators (such as LTB_4 and PGD_2) and chemokines (such as IL-8 and CC-chemokine ligand 2 (CCL2)).

of both resident cells and circulating leukocytes that are recruited to the site[4,9,35]. For example, calcitonin-gene-related peptide (CGRP), which is produced by epithelial cells, T cells, monocyte–macrophage lineage cells and possibly other sources, may contribute to the vasodilation that is associated with late-phase reactions[57].

Late-phase reactions typically develop 2–6 h after allergen exposure, and often peak after 6–9 h. It is not understood why they do not develop in all sensitized subjects, and in other patients there may be no clear clinical demarcation between the end of the early phase and the onset of the late phase[4]. In human skin, leukocytes recruited in late-phase reactions consist of T cells (T_H2 cells at early stages of the response, and T_H1 cells at late stages), which can contribute to changes in the cytokine environment at such sites), granulocytes (eosinophils and smaller numbers of neutrophils and basophils) and monocytes[58]. A similar set of cells has been found to participate in late-phase reactions that are elicited in the lower airways of patients with asthma, as determined by analysing bronchoalveolar lavage fluid[4,50,51]. Experimentally induced late-phase reactions typically resolve fully without treatment, but the mechanisms responsible largely remain to be defined.

Chronic allergic inflammation

When allergen exposure is continuous or repetitive, inflammation persists, and many innate and adaptive immune cells derived from the blood can be found in the tissues at sites of allergen challenge. This persistent inflammation is associated with changes in the structural cells at the affected sites, and in many cases with markedly altered function of the affected organs. Whereas early-phase reactions and late-phase reactions can easily be studied experimentally in human volunteers, most investigations of chronic allergic inflammation involve either experimental animal models of allergic disorders, none of which can be considered identical to the human diseases, or biopsy studies of human patients afflicted with these disorders. It is therefore not surprising that there is no clear understanding of how, after persistent and/or multiple exposures to allergen, local tissue inflammation changes from a series of early-phase and late-phase reactions to chronic allergic inflammation.

It is known that inflammation in patients with chronic asthma can involve all of the layers of the airway wall and typically is associated with: changes in the epithelium, including an increased number of goblet cells (which produce mucus); increased production of cytokines and chemokines by epithelial cells, as well as areas of epithelial injury and repair; substantial inflammation of the submucosa, including the development of increased deposition of extracellular-matrix molecules in the lamina reticularis (beneath the epithelial basement membrane); changes in fibroblasts, increased development of myofibroblasts and increased vascularity; and increased thickness of the muscular layer of the airways, with increased size, number and function of smooth muscle cells[28,59–61] (Figs 5 and 6).

Some studies[62], but not others[63], have reported increases in the number and length of tachykinin-containing nerves in the airways of patients with asthma. However, production of tachykinins by immune cells may also contribute to 'neurogenic inflammation' in asthma[64]. Patients with asthma show a marked bronchial hypersensitivity to both cholinergic and non-adrenergic, non-cholinergic (NANC) agonists of bronchoconstriction, as well as decreased sensitivity to adrenergic and NANC bronchodilators[65].

The complex interactions between affected airway epithelial cells and the underlying mesenchymal cells, which together are known as the 'epithelial–mesenchymal trophic unit' and are thought to regulate the tissue remodelling characteristic of chronic allergic inflammation of the airways, have been likened to those at a persistent wound[60]. In patients with asthma, mast cells can appear in increased numbers in the smooth muscle of the airway (Figs 5 and 6), placing this potent source of mediators that can influence smooth muscle function in intimate proximity to this crucial target-cell population[54,66]. This may contribute to the development of 'non-specific airway hyperreactivity' to agonists such as histamine, cys-LTs and methacholine, which is a hallmark of asthma[46,67].

In individuals with asthma, infections with common respiratory viruses such as rhinoviruses, influenza viruses and respiratory syncytial virus can

produce a marked exacerbation of the signs and symptoms of asthma[68]. Although the mechanisms that underlie this exacerbation are not fully understood, one factor may be the way in which the viruses affect the function of bronchial epithelial cells[68]. Mast cells appear in the airway epithelium in asthma[54] and can be activated by viral products through Toll-like receptors[34]. However, the role (if any) of mast cells in viral exacerbations of asthma remains to be determined.

In atopic dermatitis[69] and allergic rhinitis[70], as well as in asthma, chronic allergic inflammation is associated with tissue remodelling. This remodelling can involve long-term changes to the structural elements of the affected sites (such as increased vascularity) and substantial alterations in the barrier function of the affected epithelia. In many patients with allergic rhinitis, structural changes include the development of nasal polyps[70]. In atopic dermatitis, impaired function of the skin barrier is associated with a markedly increased risk of both cutaneous infections and the colonization of the affected skin with the bacterium *Staphylococcus aureus*[69]. In allergic rhinitis, impaired barrier function of the upper airway[71] may contribute to an increased susceptibility of patients to chronic sinus infections[70].

IgE and the exacerbation of allergic disorders

Many patients who initially have a single allergic disorder, such as atopic dermatitis, eventually develop others, such as allergic rhinitis and allergic asthma (this is called the allergic march or atopic march)[72]. This process may be driven in part by a vicious circle in which allergic inflammation diminishes the function of the epithelial barrier. This increases the immune system's exposure to the original allergens and additional allergens, and existing allergen-specific IgE contributes to sensitization to new allergens[21]. In this scheme, antigen-presenting cells (APCs) that express surface FcεRI and/or the low-affinity IgE receptor CD23 (including FcεRI-bearing Langerhans cells and other dendritic cells, as well as CD23-bearing B cells) capture allergens by means of their surface-bound allergen-specific IgE. By processing these IgE-bound antigens, APCs can promote the development of T_H2-cell responses to other epitopes of the allergen for which sensitization already exists or to other allergens that are being processed in parallel by the same APCs[21]. This proposed mechanism may result in epitope spreading (the production of IgE specific for multiple epitopes on single allergens and IgE specific for new allergens)[21].

In this model, the acquisition of IgE-dependent immunological reactivity to more and more allergens would occur in parallel with the clonal expansion of populations of effector T cells that can respond to any of a group of allergen-derived peptides[21]. However, a diverse range of genetic and environmental factors can influence the extent to which the pathology in individual allergic subjects depends on allergen, allergen-specific IgE, FcεRI, mast cells and basophils, as opposed to allergen-derived peptides and effector T cells (either T_H2 cells or T_H17 cells[21,73]).

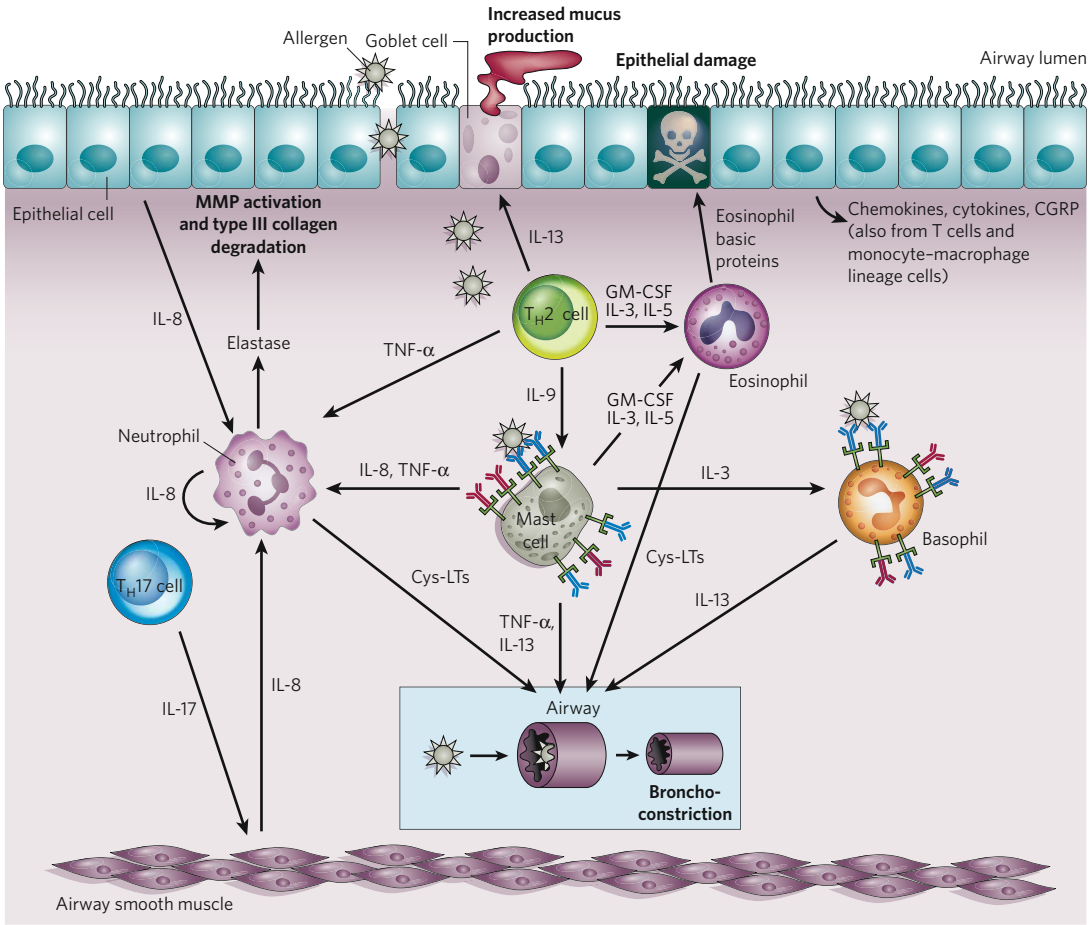


Figure 4 | Late phase of allergen-induced airway inflammation. Late-phase reactions have many features in common with early-phase reactions (Fig. 3). But late-phase reactions typically occur hours after allergen challenge and are thought to reflect the actions of innate and adaptive immune cells that have been recruited from the circulation, as well as the secretion of inflammatory mediators by tissue-resident cells. The innate immune cells include neutrophils, monocytes (not shown), eosinophils and basophils. Other cells that secrete inflammatory mediators include mast cells that have been activated by IgE- and allergen-dependent FcεRI aggregation, and tissue-resident or recruited T cells that recognize allergen-derived peptides. Therefore, in a late-phase reaction, for example, elastase released by neutrophils promotes the activation of matrix metalloproteinases (MMPs) and the degradation of type III collagen. In addition, basic proteins released by eosinophils can injure epithelial cells, and several other mediators produced by recruited or tissue-resident cells can induce bronchoconstriction. CGRP, calcitonin-gene-related peptide; GM-CSF, granulocyte–macrophage colony-stimulating factor; T_H17 cell, IL-17-producing T_H cell.

The increased levels of IgE observed in many allergic subjects can drive another amplification mechanism in allergic disorders. As local or circulating concentrations of IgE increase, mast cells and basophils display more FcεRI on their surface and have enhanced IgE-dependent effector function[8,35,74]. In addition, certain IgE molecules seem to be able to undergo antigen-independent aggregation after binding to FcεRI, thus provoking some mediator secretion by mast cells even in the absence of specific antigen[35,74]. Should this mechanism occur *in vivo*, it might contribute to the persistence of symptoms in some patients even in the absence of ongoing exposure to specific antigen.

There is strong evidence that immunoglobulin class-switch recombination can occur locally in tissues affected by allergic inflammation[21], resulting in the production of IgE. This finding can help to explain why mast cells at these sites display FcεRI molecules that remain fully saturated with IgE even when circulating IgE concentrations are relatively low[21]. It also suggests that IgE-dependent mechanisms of effector-cell activation might contribute to the development of inflammation (and related organ dysfunction) that is indistinguishable from that observed in allergic asthma, even in subjects who have low levels of IgE and in which a specific allergen has not yet been identified[21].

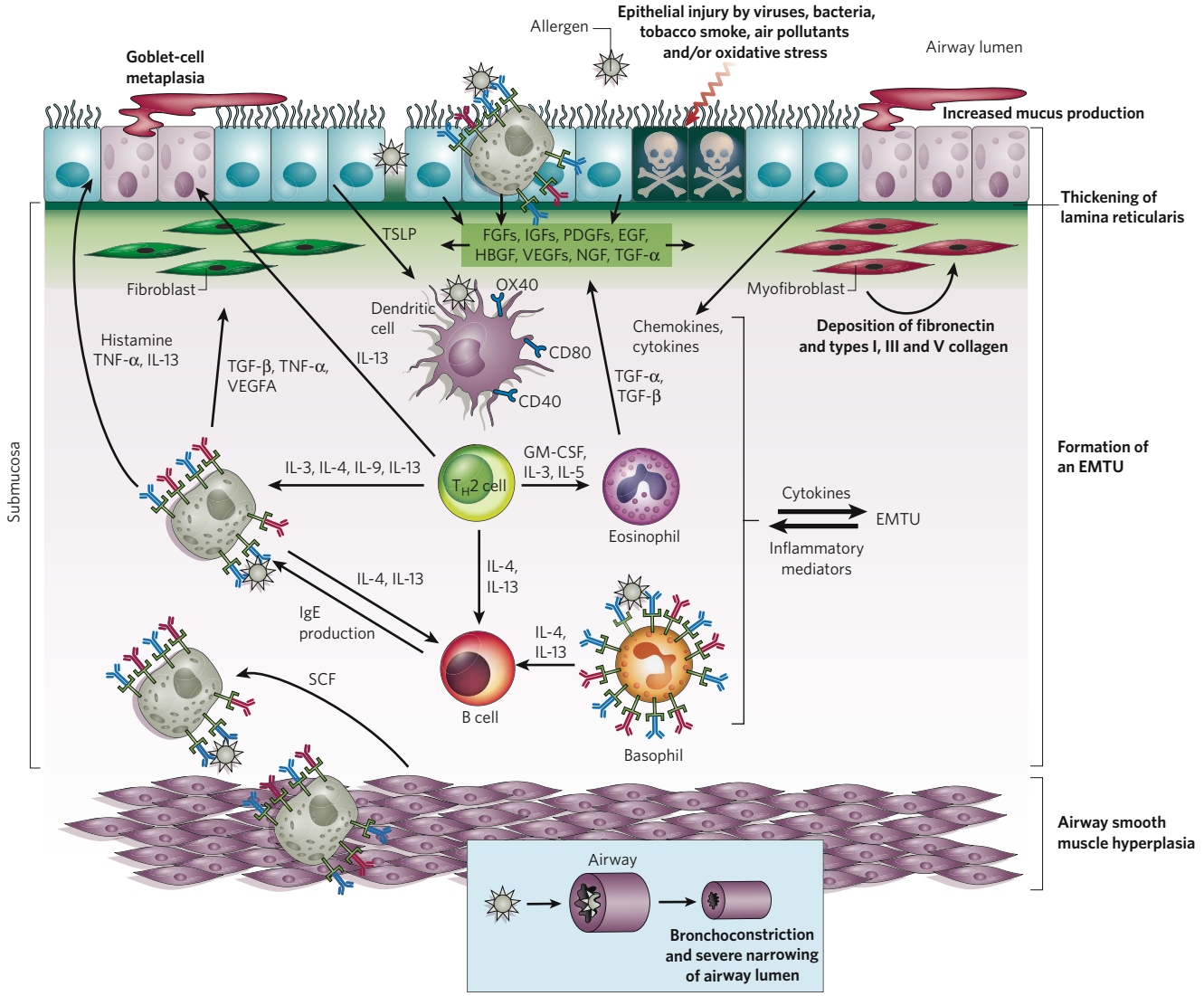


Figure 5 | Chronic stage of allergen-induced airway inflammation. In chronic allergic inflammation, repetitive or persistent exposure to allergens has several effects. Innate immune cells (including eosinophils, basophils, neutrophils and monocyte–macrophage lineage cells) and adaptive immune cells (including T_H2 cells, other types of T cells, and B cells) take up residence in the tissues. In addition, more mast cells develop in the tissue, and these cells display large amounts of IgE bound to FcεRI and have an altered anatomical distribution. Last, complex interactions are initiated between recruited and tissue-resident innate and adaptive immune cells, epithelial cells and structural cells (such as fibroblasts, myofibroblasts and airway smooth muscle cells) and blood vessels and lymphatic vessels, and nerves (not shown). Repetitive epithelial injury due to chronic allergic inflammation can be exacerbated by exposure to pathogens or environmental factors, and the consequent repair response results in an epithelial–mesenchymal trophic unit (EMTU) being established. This unit is thought to sustain T_H2-cell-associated inflammation, to promote sensitization to additional allergens or allergen epitopes (for example, epithelial-cell-derived TSLP can upregulate the expression of co-stimulatory molecules such as OX40, CD40 and CD80 by dendritic cells), and to regulate the airway remodelling process. These processes result in many functionally important changes in the structure of the affected tissue. These changes include substantial thickening of the airway walls (including the epithelium, lamina reticularis, submucosa and smooth muscle), increased deposition of extracellular-matrix proteins (such as fibronectin, and type I, III and V collagen), and hyperplasia of goblet cells, which is associated with increased mucus production. In individuals who have such thickened airway walls, bronchoconstriction can result in more severe narrowing of the airway lumen than occurs in airways with normal wall thickness. In some individuals, especially those with severe asthma, T_H17 cells (which secrete IL-17) may also contribute to the recruitment of neutrophils to sites of inflammation (not shown). EGF, epidermal growth factor; FGF, fibroblast growth factor; HBEGF, heparin-binding EGF-like growth factor; IGF, insulin-like growth factor; NGF, nerve growth factor; PDGF, platelet-derived growth factor; SCF, stem-cell factor (also known as KIT ligand).

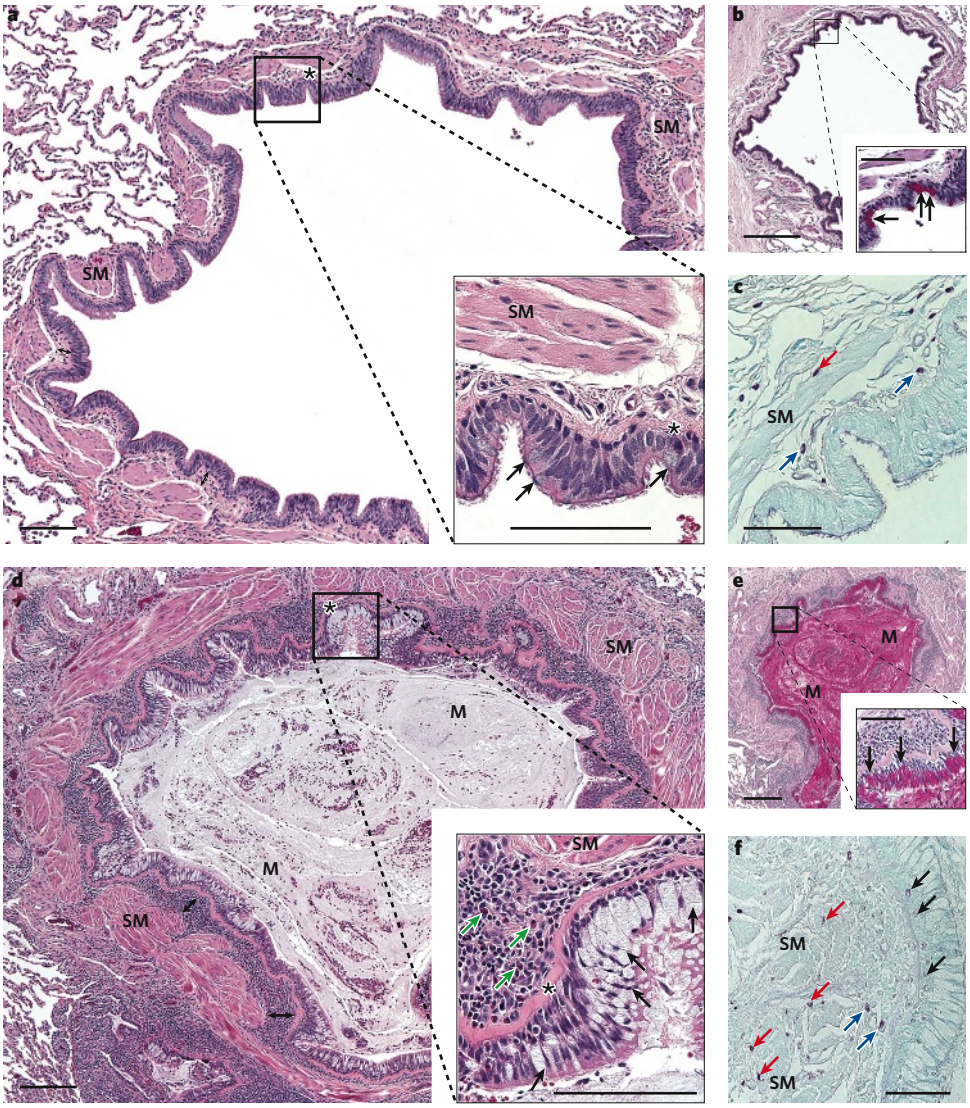


Figure 6 | Chronic allergic inflammation and tissue remodelling in asthma. Tissue sections from the airway of a non-asthmatic person (**a**–**c**) and a patient with severe asthma (**d**–**f**) are shown. Specimens were taken from lung resections (carried out for other indications), fixed in 10% neutral buffered formalin and processed routinely; sections 5 μm thick, from the same area of tissue, were stained with haematoxylin and eosin (**a** and **d**), periodic acid–Schiff with diastase (to stain mucus red; **b** and **e**), or pinacyanol erythrosinate (to stain mast cells purple; **c** and **f**). Scale bars, 500 μm (**a** and **d**), 100 μm (inset **a** and **d**), 400 μm (**b** and **e**), 100 μm (inset **b** and **e**) and 100 μm (**c** and **f**). **a**–**c**, A normal small bronchus. There are few goblet cells (black arrows in insets) in the epithelium. The basement membrane and underlying lamina reticularis (at asterisk in **a**, hardly visible at this magnification) are normal. The submucosa (the length of the double-headed arrows in **a**) contains few leukocytes and the occasional mast cell (blue arrows in **c**), and the bronchial smooth muscle (SM) has few adjacent mast cells (red arrow in **c**). **d**–**f**, A small bronchus from a patient with a history of severe asthma. Mucus (M) fills the airway lumen (**d** and **e**). There are many goblet cells (black arrows in insets) and the occasional intra-epithelial mast cell (black arrows in **f**). The lamina reticularis (asterisk in inset in **d**) is markedly thickened. The submucosa (double-headed arrows in **d**) contains many eosinophils (green arrows in inset in **d**) and other leukocytes, as well as mast cells (blue arrows in **f**). There is more bronchial smooth muscle (SM) than in **a**–**c**, and there are many mast cells (red arrows in **f**) among bundles of smooth muscle cells. (Figure courtesy of G. J. Berry, Stanford University, California.)

In addition, several effector mechanisms that are independent of IgE may also contribute to the pathology of allergic inflammation. In a mouse model of chronic asthma, mast cells can substantially influence features of chronic allergic inflammation and tissue remodelling (including expansion of the number of goblet cells), independently of mast-cell signalling through either IgE–FcεRI or antigen–IgG_1–FcγRIII[75]. Thus mast cells have the potential to drive important features of allergic inflammation independently of IgE.

Moreover, in mouse models, allergic inflammation of the airways can be induced in mice that lack mast cells or B cells[75]. This underscores the important point that the coordination of chronic allergic inflammation may reflect complex and partly redundant pathways involving interactions between mast cells[35,51,52,55,75], T cells[4,9,56], eosinophils[76], basophils[50,51], neutrophils[73], monocyte–macrophage lineage cells[77], platelets[78] and natural killer T cells[79,80], as well as a large and growing list of cytokines (including IL-4, IL-5, IL-12, IL-13, IL-15, IL-25 and IL-33). However, the relative importance of each of these potential effector or regulatory elements may vary in different disorders or between patients, and many of these interactions may not be markedly affected by IgE. This possibility may explain, at least in part, why the humanized IgE-specific monoclonal antibody known as omalizumab (Xolair) has shown variable clinical effectiveness in patients with moderate-to-severe asthma[81]. Indeed, the results of attempts to target IgE-dependent mechanisms of inflammation in various allergic disorders support many other lines of evidence indicating that IgE has an important pathological role in some subjects with moderate-to-severe asthma[81], allergic rhinitis[81,82] or certain food allergies[83], whereas T-cell-dependent effector mechanisms are more important in most patients with atopic dermatitis and perhaps in some with asthma as well[4,58,69].

Suppression and resolution of allergic inflammation

Apart from the cessation of allergen-specific stimulation of effector cells, as occurs at the end of the pollen season in pollen-sensitive individuals, the factors that regulate the resolution of allergic inflammation are poorly understood. Some effector cells may undergo apoptosis as concentrations of cytokines that promote the survival of such cells locally diminish[84]; others (such as mast cells) may decrease the extent to which they differentiate, mature or proliferate locally[85]; and others may emigrate from the affected site[86].

In some models of allergic contact hypersensitivity, the production of IL-10 by mast cells contributes significantly to the ability of mast cells to reduce many features of inflammation in the affected sites[87]. Whether similar anti-inflammatory or immunosuppressive actions of mast cells can be elicited in the context of IgE-associated allergic inflammation remains to be determined. However, several types of innate and adaptive immune cells that infiltrate sites of allergic inflammation (including eosinophils and various populations of regulatory T cells) can produce mediators, cytokines, chemokines and growth factors that could reduce inflammation or promote repair at these sites. Such products include the resolvin and protectin lipid mediators[88], IL-4 (which can have anti-inflammatory effects[89]), TGF-β[90,91], TGF-α[92], IL-10 (refs 16, 87, 89, 91, 93) and IL-35 (ref. 93).

Allergen-specific regulatory T cells have been reported in patients after allergen-specific immunotherapy[10,11,16,93]. In addition, there is evidence from animal models of allergy and asthma that both antigen-specific regulatory T cells and naturally occurring regulatory T cells can limit disease, in part by IL-10- and TGF-β-dependent mechanisms[16,93]. However, the extent to which particular populations of regulatory T cells can limit allergic inflammation at the times of exposure to specific allergen, or help to resolve allergic inflammation when exposure to allergen ceases, and the mechanisms by which the regulatory T cells exert these effects remain to be fully understood.

Management of allergies and allergic inflammation

The two key elements of allergy management are preventing the exposure of sensitized individuals to allergen and treating these individuals with therapeutic agents appropriate to the disorder. For example, antihistamines that target the H_1 histamine receptor are a mainstay of treatment for allergic rhinitis but have been of limited value in asthma[7,9]. Asthma is generally treated with inhaled corticosteroids (which suppress many of the pathways that contribute to inflammation) and agonists of β-adrenergic receptors (which induce bronchodilation). These treatments are effective in many (but not all) subjects[7,9]. Some patients with asthma are helped by drugs that target cys-LTs[7,9]. Omalizumab, which targets IgE, helps some subjects with moderate or severe asthma[81] and is being evaluated in other settings[82].

The extent to which pharmacogenetic approaches can be used to understand the basis of variable clinical responses to the same agent, and to identify subjects who will benefit from particular treatments, is an area of active investigation[94]. Allergen-specific immunotherapy should be considered in situations in which this approach has been shown to be beneficial[10,11].

Many new pharmacological or biological agents that target the various steps in the cell and mediator pathways implicated in allergic inflammation are being investigated[7,9]. Some of these compounds are designed to exploit endogenous mechanisms to suppress effector-cell activation during allergic inflammation, such as co-engagement of FcεRI with the inhibitory receptor FcγRIIB[95], or to take advantage of other mechanisms that can negatively regulate FcεRI-dependent signalling[8,36].

Strategies to reduce sensitization and promote tolerance to common allergens are also being considered. One example is reducing exposure to allergens through routes that favour the generation of T_H2-cell responses (for example, the skin and respiratory tract) while increasing exposure through routes that favour the production of tolerance (the gastrointestinal tract)[96]. Other approaches include attempting to devise vaccines that can be used (carefully) to induce tolerance to substances before natural sensitization can occur[7,9–11], using probiotics to promote the development of a 'healthy' immune system (that is, biased to T_H1-cell responses or modified T_H2-cell responses) [12], and using data derived from epidemiological studies to promote aspects of lifestyle that may reduce the risk of developing allergic disorders. Examples of this last approach include reducing exposure to common aeroallergens[97], increasing exposure to certain pets (such as dogs)[98], and increasing exercise and outdoor activities[99].

What next?

The recent progress in our understanding of the genetic, environmental, tissue-specific and immunological factors that contribute to the development of allergic disorders and allergic inflammation has suggested possible new approaches for managing, treating or even preventing these disorders. Will more specific or potent targeting of additional mediators, their receptors, IgE, FcεRI or effector cells (such as mast cells, T cells and natural killer T cells), used alone or in combination, afford a substantial improvement over current approaches? Will marshalling the current knowledge of the immunobiology of allergy and tolerance allow researchers to devise ways to prevent allergic sensitization (for example, by improving epithelial barrier function in individuals in whom it is impaired) or to induce tolerance by safer and more effective forms of allergen-specific immunotherapy? Time will tell. Such efforts are important because, although most patients with allergic disorders can be helped by current management strategies, these complex 'disorders of advanced civilization' have so far been difficult to control in many patients, let alone to prevent or cure. ■

1. von Pirquet, C. Allergie. *Münch. Med. Wochenschr.* **53,** 1457–1458 (1906).
2. Silverstein, A. M. Clemens Freiherr von Pirquet: explaining immune complex disease in 1906. *Nature Immunol.* **1,** 453–455 (2000).
3. Holgate, S. T. The epidemic of allergy and asthma. *Nature* **402,** B2–B4 (1999).
4. Kay, A. B. Allergy and allergic diseases. First of two parts. *N. Engl. J. Med.* **344,** 30–37 (2001).
5. Eder, W., Ege, M. J. & von Mutius, E. The asthma epidemic. *N. Engl. J. Med.* **355,** 2226–2235 (2006).
6. Sampson, H. A. *et al.* Symposium on the definition and management of anaphylaxis: summary report. *J. Allergy Clin. Immunol.* **115,** 584–591 (2005).
7. Barnes, P. J. New therapies for asthma. *Trends Mol. Med.* **12,** 515–520 (2006).
8. Kraft, S. & Kinet, J. P. New developments in FcεRI regulation, function and inhibition. *Nature Rev. Immunol.* **7,** 365–378 (2007).
   This review of FcεRI-dependent signalling in mast cells and basophils considers how the biology and functional properties of FcεRI might be exploited for the development of new therapeutics.
9. Holgate, S. T. & Polosa, R. Treatment strategies for allergy and asthma. *Nature Rev. Immunol.* **8,** 218–230 (2008).
10. Larché, M., Akdis, C. A. & Valenta, R. Immunological mechanisms of allergen-specific immunotherapy. *Nature Rev. Immunol.* **6,** 761–771 (2006).
    This review discusses the history, immunological mechanisms and future prospects of improving allergen-specific immunotherapy.
11. Akdis, M. & Akdis, C. A. Mechanisms of allergen-specific immunotherapy. *J. Allergy Clin. Immunol.* **119,** 780–791 (2007).
12. Kukkonen, K. *et al.* Probiotics and prebiotic galacto-oligosaccharides in the prevention of allergic diseases: a randomized, double-blind, placebo-controlled trial. *J. Allergy Clin. Immunol.* **119,** 192–198 (2007).
13. Yazdanbakhsh, M., Kremsner, P. G. & van Ree, R. Allergy, parasites, and the hygiene hypothesis. *Science* **296,** 490–494 (2002).
14. Galli, S. J. & Askenase, P. W. in *The Reticuloendothelial System: A Comprehensive Treatise* Vol. IX: Hypersensitivity (eds Abramoff, P., Phillips, S. M. & Escobar, M. R.) 321–369 (Plenum, 1986).
15. Fallon, P. G. & Mangan, N. E. Suppression of T_H2-type allergic reactions by helminth infection. *Nature Rev. Immunol.* **7,** 220–230 (2007).
16. Hawrylowicz, C. M. & O'Garra, A. Potential role of interleukin-10-secreting regulatory T cells in allergy and asthma. *Nature Rev. Immunol.* **5,** 271–283 (2005).
    This review provides an introduction to the mechanisms by which regulatory T cells that produce the anti-inflammatory and immunosuppressive cytokine IL-10 might limit the pathology associated with allergy and allergic inflammation of the airways in asthma.
17. Romagnani, S. Coming back to a missing immune deviation as the main explanatory mechanism for the hygiene hypothesis. *J. Allergy Clin. Immunol.* **119,** 1511–1513 (2007).
18. Cookson, W. The immunogenetics of asthma and eczema: a new focus on the epithelium. *Nature Rev. Immunol.* **4,** 978–988 (2004).
19. Vercelli, D. Discovering susceptibility genes for asthma and allergy. *Nature Rev. Immunol.* **8,** 169–182 (2008).
    This review presents the current understanding of the many genes that have been implicated in asthma and allergy, including evidence that exposure to the same microbial products may have opposite effects on susceptibility to developing allergic disorders, depending on an individual's genotype.
20. Geha, R. S., Jabara, H. H. & Brodeur, S. R. The regulation of immunoglobulin E class-switch recombination. *Nature Rev. Immunol.* **3,** 721–732 (2003).
21. Gould, H. J. & Sutton, B. J. IgE in allergy and asthma today. *Nature Rev. Immunol.* **8,** 205–217 (2008).
    This review describes the complex role of IgE and its receptors in allergy and asthma, including evidence that IgE and its receptors may contribute to epitope spreading in, and therefore exacerbation of, allergic disorders.
22. Herrick, C. A. & Bottomly, K. To respond or not to respond: T cells in allergic asthma. *Nature Rev. Immunol.* **3,** 405–412 (2003).
23. Dickey, B. F. Exoskeletons and exhalation. *N. Engl. J. Med.* **357,** 2082–2084 (2007).
24. Saxon, A. & Diaz-Sanchez, D. Air pollution and allergy: you are what you breathe. *Nature Immunol.* **6,** 223–226 (2005).
25. Hammad, H. & Lambrecht, B. N. Dendritic cells and epithelial cells: linking innate and adaptive immunity in asthma. *Nature Rev. Immunol.* **8,** 193–204 (2008).
26. Platts-Mills, T. A., Woodfolk, J. A., Erwin, E. A. & Aalberse, R. Mechanisms of tolerance to inhalant allergens: the relevance of a modified T_H2 response to allergens from domestic animals. *Springer Semin. Immunopathol.* **25,** 271–279 (2004).
27. Sandilands, A., Smith, F. J., Irvine, A. D. & McLean, W. H. Filaggrin's fuller figure: a glimpse into the genetic architecture of atopic dermatitis. *J. Invest. Dermatol.* **127,** 1282–1284 (2007).
28. Schleimer, R. P., Kato, A., Kern, R., Kuperman, D. & Avila, P. C. Epithelium: at the interface of innate and adaptive immune responses. *J. Allergy Clin. Immunol.* **120,** 1279–1284 (2007).
29. Ying, S., Meng, Q., Corrigan, C. J. & Lee, T. H. Lack of filaggrin expression in the human bronchial mucosa. *J. Allergy Clin. Immunol.* **118,** 1386–1388 (2006).
30. Jeong, S. K. *et al.* Mite and cockroach allergens activate protease-activated receptor 2 and delay epidermal permeability barrier recovery. *J. Invest. Dermatol.* **128,** 1930–1939 (2008).
31. McKerrow, J. H., Caffrey, C., Kelly, B., Loke, P. & Sajid, M. Proteases in parasitic diseases. *Annu. Rev. Pathol.* **1,** 497–536 (2006).
32. Min, B. & Paul, W. E. Basophils: in the spotlight at last. *Nature Immunol.* **9,** 223–225 (2008).
33. Sokol, C. L., Barton, G. M., Farr, A. G. & Medzhitov, R. A mechanism for the initiation of allergen-induced T helper type 2 responses. *Nature Immunol.* **9,** 310–318 (2008).
    This paper identifies a key role for basophils in the initiation of T_H2-cell responses to exogenous proteases.
34. Marshall, J. S. Mast-cell responses to pathogens. *Nature Rev. Immunol.* **4,** 787–799 (2004).
35. Galli, S. J. *et al.* Mast cells as 'tunable' effector and immunoregulatory cells: recent advances. *Annu. Rev. Immunol.* **23,** 749–786 (2005).

This review discusses many aspects of mast-cell biology, including mast-cell phenotypic heterogeneity and function, and the roles of mast cells as effector and potential immunoregulatory cells in innate and adaptive immune responses.
36. Gilfillan, A. M. & Tkaczyk, C. Integrated signalling pathways for mast-cell activation. *Nature Rev. Immunol.* **6,** 218–230 (2006).
37. Rivera, J. & Gilfillan, A. M. Molecular regulation of mast cell activation. *J. Allergy Clin. Immunol.* **117,** 1214–1225 (2006).
38. Dvorak, A. M. Ultrastructural studies of human basophils and mast cells. *J. Histochem. Cytochem.* **53,** 1043–1070 (2005).
39. Caughey, G. H. Mast cell tryptases and chymases in inflammation and host defense. *Immunol. Rev.* **217,** 141–154 (2007).
40. Pejler, G., Abrink, M., Ringvall, M. & Wernersson, S. Mast cell proteases. *Adv. Immunol.* **95,** 167–255 (2007).
41. Stevens, R. L. & Adachi, R. Protease-proteoglycan complexes of mouse and human mast cells and importance of their beta-tryptase-heparin complexes in inflammation and innate immunity. *Immunol. Rev.* **217,** 155–167 (2007).
42. Bradding, P. & Holgate, S. T. The mast cell as a source of cytokines in asthma. *Ann. NY Acad. Sci.* **796,** 272–281 (1996).
43. Saito, H., Nakajima, T. & Matsumoto, K. Human mast cell transcriptome project. *Int. Arch. Allergy Immunol.* **125,** 1–8 (2001).
44. Boyce, J. A. Mast cells and eicosanoid mediators: a system of reciprocal paracrine and autocrine regulation. *Immunol. Rev.* **217,** 168–185 (2007).
45. Finkelman, F. D. Anaphylaxis: lessons from mouse models. *J. Allergy Clin. Immunol.* **120,** 506–515 (2007).
46. Wills-Karp, M. Immunologic basis of antigen-induced airway hyperresponsiveness. *Annu. Rev. Immunol.* **17,** 255–281 (1999).
47. Sarin, S., Undem, B., Sanico, A. & Togias, A. The role of the nervous system in rhinitis. *J. Allergy Clin. Immunol.* **118,** 999–1016 (2006).
48. Cevikbas, F., Steinhoff, A., Homey, B. & Steinhoff, M. Neuroimmune interactions in allergic skin diseases. *Curr. Opin. Allergy Clin. Immunol.* **7,** 365–373 (2007).
49. Lalloo, U. G., Barnes, P. J. & Chung, K. F. Pathophysiology and clinical presentations of cough. *J. Allergy Clin. Immunol.* **98,** S91–S96; discussion S96–S97 (1996).
50. MacGlashan, D. Jr, Gauvreau, G. & Schroeder, J. T. Basophils in airway disease. *Curr. Allergy Asthma Rep.* **2,** 126–132 (2002).
51. Marone, G., Triggiani, M. & de Paulis, A. Mast cells and basophils: friends as well as foes in bronchial asthma? *Trends Immunol.* **26,** 25–31 (2005).
52. Galli, S. J., Grimbaldeston, M. A. & Tsai, M. Immunomodulatory mast cells: negative, as well as positive, regulators of immunity. *Nature Rev. Immunol.* **8,** 478–486 (2008).
53. Sayed, B. A., Christy, A., Quirion, M. R. & Brown, M. A. The master switch: the role of mast cells in autoimmunity and tolerance. *Annu. Rev. Immunol.* **26,** 705–739 (2008).
54. Bradding, P., Walls, A. F. & Holgate, S. T. The role of the mast cell in the pathophysiology of asthma. *J. Allergy Clin. Immunol.* **117,** 1277–1284 (2006).
55. Brown, J. M., Wilson, T. M. & Metcalfe, D. D. The mast cell and allergic diseases: role in pathogenesis and implications for therapy. *Clin. Exp. Allergy* **38,** 4–18 (2008).
56. Larché, M., Robinson, D. S. & Kay, A. B. The role of T lymphocytes in the pathogenesis of asthma. *J. Allergy Clin. Immunol.* **111,** 450–463 (2003).
57. Kay, A. B. *et al.* Airway expression of calcitonin gene-related peptide in T-cell peptide-induced late asthmatic reactions in atopics. *Allergy* **62,** 495–503 (2007).
58. Bonness, S. & Bieber, T. Molecular basis of atopic dermatitis. *Curr. Opin. Allergy Clin. Immunol.* **7,** 382–386 (2007).
59. Doherty, T. & Broide, D. Cytokines and growth factors in airway remodeling in asthma. *Curr. Opin. Immunol.* **19,** 676–680 (2007).
60. Holgate, S. T. Epithelium dysfunction in asthma. *J. Allergy Clin. Immunol.* **120,** 1233–1244 (2007).
This review discusses the role of the airway epithelium and its function (and dysfunction) in the development and pathology of asthma.
61. Mauad, T., Bel, E. H. & Sterk, P. J. Asthma therapy and airway remodeling. *J. Allergy Clin. Immunol.* **120,** 997–1009 (2007).
62. Ollerenshaw, S. L., Jarvis, D., Sullivan, C. E. & Woolcock, A. J. Substance P immunoreactive nerves in airways from asthmatics and nonasthmatics. *Eur. Respir. J.* **4,** 673–682 (1991).
63. Chanez, P. *et al.* Bronchial mucosal immunoreactivity of sensory neuropeptides in severe airway diseases. *Am. J. Respir. Crit. Care Med.* **158,** 985–990 (1998).
64. Joos, G. F., De Swert, K. O., Schelfhout, V. & Pauwels, R. A. The role of neural inflammation in asthma and chronic obstructive pulmonary disease. *Ann. NY Acad. Sci.* **992,** 218–230 (2003).
65. Lewis, M. J., Short, A. L. & Lewis, K. E. Autonomic nervous system control of the cardiovascular and respiratory systems in asthma. *Respir. Med.* **100,** 1688–1705 (2006).
66. Brightling, C. E. *et al.* Mast-cell infiltration of airway smooth muscle in asthma. *N. Engl. J. Med.* **346,** 1699–1705 (2002).
67. Cohn, L., Elias, J. A. & Chupp, G. L. Asthma: mechanisms of disease persistence and progression. *Annu. Rev. Immunol.* **22,** 789–815 (2004).
68. Gern, J. E. & Busse, W. W. Relationship of viral infections to wheezing illnesses and asthma. *Nature Rev. Immunol.* **2,** 132–138 (2002).
69. Leung, D. Y., Boguniewicz, M., Howell, M. D., Nomura, I. & Hamid, Q. A. New insights into atopic dermatitis. *J. Clin. Invest.* **113,** 651–657 (2004).
70. Pawankar, R., Nonaka, M., Yamagishi, S. & Yagi, T. Pathophysiologic mechanisms of chronic rhinosinusitis. *Immunol. Allergy Clin. North Am.* **24,** 75–85 (2004).
71. Takano, K. *et al.* HLA-DR- and CD11c-positive dendritic cells penetrate beyond well-developed epithelial tight junctions in human nasal mucosa of allergic rhinitis. *J. Histochem. Cytochem.* **53,** 611–619 (2005).
72. Spergel, J. M. & Paller, A. S. Atopic dermatitis and the atopic march. *J. Allergy Clin. Immunol.* **112,** S118–S127 (2003).
73. Barnes, P. J. Immunology of asthma and chronic obstructive pulmonary disease. *Nature Rev. Immunol.* **8,** 183–192 (2008).
74. Kawakami, T. & Galli, S. J. Regulation of mast-cell and basophil function and survival by IgE. *Nature Rev. Immunol.* **2,** 773–786 (2002).
75. Yu, M. *et al.* Mast cells can promote the development of multiple features of chronic asthma in mice. *J. Clin. Invest.* **116,** 1633–1641 (2006).
This paper presents evidence that mast cells can contribute to multiple features of the pathology in a mouse model of chronic asthma both by mechanisms that do or do not require the antibody (IgE and/or IgG1)-dependent activation of mast cells through the FcRγ chain shared by mast-cell FcεRI and FcγRIII.
76. Jacobsen, E. A. *et al.* Allergic pulmonary inflammation in mice is dependent on eosinophil-induced recruitment of effector T cells. *J. Exp. Med.* **205,** 699–710 (2008).
77. Peters-Golden, M. The alveolar macrophage: the forgotten cell in asthma. *Am. J. Respir. Cell Mol. Biol.* **31,** 3–7 (2004).
78. Kasperska-Zajac, A. & Rogala, B. Platelet activation during allergic inflammation. *Inflammation* **30,** 161–166 (2007).
79. Akbari, O. *et al.* $CD4^+$ invariant T-cell-$receptor^+$ natural killer T cells in bronchial asthma. *N. Engl. J. Med.* **354,** 1117–1129 (2006).
80. Vijayanand, P. *et al.* Invariant natural killer T cells in asthma and chronic obstructive pulmonary disease. *N. Engl. J. Med.* **356,** 1410–1422 (2007).
81. Holgate, S. T., Djukanovic, R., Casale, T. & Bousquet, J. Anti-immunoglobulin E treatment with omalizumab in allergic diseases: an update on anti-inflammatory activity and clinical efficacy. *Clin. Exp. Allergy* **35,** 408–416 (2005).
82. Casale, T. B. *et al.* Effect of omalizumab on symptoms of seasonal allergic rhinitis: a randomized controlled trial. *J. Am. Med. Assoc.* **286,** 2956–2967 (2001).
83. Leung, D. Y. *et al.* Effect of anti-IgE therapy in patients with peanut allergy. *N. Engl. J. Med.* **348,** 986–993 (2003).
84. Akdis, C. A., Blaser, K. & Akdis, M. Apoptosis in tissue inflammation and allergic disease. *Curr. Opin. Immunol.* **16,** 717–723 (2004).
85. Ryan, J. J. *et al.* Mast cell homeostasis: a fundamental aspect of allergic disease. *Crit. Rev. Immunol.* **27,** 15–32 (2007).
86. Medoff, B. D., Thomas, S. Y. & Luster, A. D. T cell trafficking in allergic asthma: the ins and outs. *Annu. Rev. Immunol.* **26,** 205–232 (2008).
This is a comprehensive review of the molecular regulation and consequences of T-cell migration in allergic inflammation of airways.
87. Grimbaldeston, M. A., Nakae, S., Kalesnikoff, J., Tsai, M. & Galli, S. J. Mast cell-derived interleukin 10 limits skin pathology in contact dermatitis and chronic irradiation with ultraviolet B. *Nature Immunol.* **8,** 1095–1104 (2007).
88. Serhan, C. N., Yacoubian, S. & Yang, R. Anti-inflammatory and proresolving lipid mediators. *Annu. Rev. Pathol.* **3,** 279–312 (2008).
89. Opal, S. M. & DePalo, V. A. Anti-inflammatory cytokines. *Chest* **117,** 1162–1172 (2000).
90. Letterio, J. J. & Roberts, A. B. Regulation of immune responses by TGF-β. *Annu. Rev. Immunol.* **16,** 137–161 (1998).
91. Li, M. O. & Flavell, R. A. Contextual regulation of inflammation: a duet by transforming growth factor-β and interleukin-10. *Immunity* **28,** 468–476 (2008).
92. Burgel, P. R. *et al.* Human eosinophils induce mucin production in airway epithelial cells via epidermal growth factor receptor activation. *J. Immunol.* **167,** 5948–5954 (2001).
93. Vignali, D. A. A., Collison, L. W. & Workman, C. J. How regulatory T cells work. *Nature Rev. Immunol.* **8,** 523–532 (2008).
94. Hall, I. P. & Sayers, I. Pharmacogenetics and asthma: false hope or new dawn? *Eur. Respir. J.* **29,** 1239–1245 (2007).
95. Zhu, D. *et al.* A chimeric human–cat fusion protein blocks cat-induced allergy. *Nature Med.* **11,** 446–449 (2005).
96. Lack, G., Fox, D., Northstone, K. & Golding, J. Factors associated with the development of peanut allergy in childhood. *N. Engl. J. Med.* **348,** 977–985 (2003).
97. Platts-Mills, T. A., Vervloet, D., Thomas, W. R., Aalberse, R. C. & Chapman, M. D. Indoor allergens and asthma: report of the Third International Workshop. *J. Allergy Clin. Immunol.* **100,** S2–S24 (1997).
98. Ownby, D. R. & Johnson, C. C. Does exposure to dogs and cats in the first year of life influence the development of allergic sensitization? *Curr. Opin. Allergy Clin. Immunol.* **3,** 517–522 (2003).
99. Lucas, S. R. & Platts-Mills, T. A. Physical activity and exercise in asthma: relevance to etiology and treatment. *J. Allergy Clin. Immunol.* **115,** 928–934 (2005).

Acknowledgements We thank G. Berry and J. Kalesnikoff for help with the figures, C. M. Hawrylowicz and members of the Galli laboratory for critical reading of the manuscript, and the National Institutes of Health for financial support.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Correspondence should be addressed to S.J.G. (sgalli@stanford.edu).

From endoplasmic-reticulum stress to the inflammatory response

Kezhong Zhang[1,2] & Randal J. Kaufman[1,3,4]

The endoplasmic reticulum is responsible for much of a cell's protein synthesis and folding, but it also has an important role in sensing cellular stress. Recently, it has been shown that the endoplasmic reticulum mediates a specific set of intracellular signalling pathways in response to the accumulation of unfolded or misfolded proteins, and these pathways are collectively known as the unfolded-protein response. New observations suggest that the unfolded-protein response can initiate inflammation, and the coupling of these responses in specialized cells and tissues is now thought to be fundamental in the pathogenesis of inflammatory diseases. The knowledge gained from this emerging field will aid in the development of therapies for modulating cellular stress and inflammation.

Inflammation is the first response of the immune system to infection or tissue injury, leading to protection of the human body against these insults. But prolonged or chronic inflammation is detrimental and has an important role in the development of diseases such as arthritis, Alzheimer's disease, type 1 and type 2 diabetes and cardiovascular disease[1–3].

An inflammatory response begins when cells of the immune system and/or cells involved in metabolic pathways sense pathogens, irritants and cellular damage, triggering the release of inflammatory substances, including cytokines, free radicals, hormones and other small molecules. These inflammatory substances further stimulate the cells that secreted them and target specialized cells in immune and metabolic pathways, thereby altering cellular physiology to contribute to wound healing and pathogen resistance[1]. However, there is epidemiological, clinical and experimental evidence that cellular stress (that is, impaired biological processes within the cell) and excessive inflammation are causally linked to various metabolic conditions, such as obesity, type 1 and type 2 diabetes and atherosclerosis[2,3].

Through recent intensive efforts, knowledge of the cellular and molecular mechanisms that control the inflammatory response has rapidly grown. However, crucial questions about how an inflammatory response originates have yet to be answered. For example, how does a cell interpret the presence of extracellular insults or metabolic overload and start transmitting signals that trigger an inflammatory response? Does the signalling in stress responses and inflammatory responses stem from a common mechanism or from different mechanisms that subsequently become integrated?

Recently, a set of intracellular pathways that signal the presence of cellular stress was identified. These pathways are collectively known as the unfolded-protein response (UPR), and studies of the UPR have broadened the understanding of the mechanisms by which inflammation can be initiated. Here we describe the research that has defined the molecular and cellular underpinnings of UPR-associated inflammation and then discuss how the UPR is coupled to inflammation in health and disease.

ER stress and the UPR in mammals

The endoplasmic reticulum (ER) is a membranous network of branching tubules and flattened sacs that is present in all eukaryotic cells. It extends throughout the cytoplasm of the cell and is contiguous with the nuclear envelope. The ER is mainly recognized as a protein-folding factory, responsible for the biosynthesis, folding, assembly and modification of numerous soluble proteins and membrane proteins[4]. About one-third of newly synthesized proteins translocate to the lumen of the ER, where they are folded into the correct three-dimensional structures before being targeted to various cellular organelles or transported to the surface of the cell. The ER also functions as a dynamic calcium store, which responds to growth factors, hormones, and stimuli that perturb cellular energy levels, nutrient availability or redox status. The ER seems to be a key site where intracellular signals mediated by these factors are sensed, integrated and transmitted, allowing the coordination of downstream responses. Physiological states that increase the demand for protein folding, or stimuli that disrupt the reactions by which proteins fold, create an imbalance between the protein-folding load and the capacity of the ER, causing unfolded or misfolded proteins to accumulate in the ER lumen — a condition referred to as ER stress. To ensure the fidelity of protein folding and to prevent such an accumulation of unfolded or misfolded proteins, eukaryotic cells have evolved the UPR, which alters a cell's transcriptional and translational programs to cope with stressful conditions and to resolve the protein-folding defect[5,6].

In mammalian cells, the main UPR signalling cascades are initiated by three ER-localized protein sensors: IRE1α (inositol-requiring 1α), PERK (double-stranded RNA-dependent protein kinase (PKR)-like ER kinase) and ATF6 (activating transcription factor 6)[5,6]. Each of these transmembrane proteins has an ER-luminal domain that senses unfolded proteins, a transmembrane domain by which it is targeted to the ER membrane, and a cytosolic domain that transmits signals to the transcriptional or translational apparatus. IRE1α has protein-kinase activity and site-specific endoribonuclease (RNase) activity (the functions of which are described later)[7,8]. PERK also has protein-kinase activity and functions to

[1]Department of Biological Chemistry, The University of Michigan Medical Center, 1150 West Medical Center Drive, Ann Arbor, Michigan 48109, USA. [2]Present address: Center for Molecular Medicine and Genetics, Wayne State University School of Medicine, 540 East Canfield Avenue, Detroit, Michigan 48201, USA. [3]Department of Internal Medicine, The University of Michigan Medical Center, 1150 West Medical Center Drive, Ann Arbor, Michigan 48109, USA. [4]Howard Hughes Medical Institute, The University of Michigan Medical Center, 1150 West Medical Center Drive, Ann Arbor, Michigan 48109, USA.

phosphorylate the α-subunit of eukaryotic translation-initiation factor 2α (eIF2α)[9,10]. ATF6 is a bZIP (basic region and leucine zipper)-domain-containing transcription factor belonging to the CREB (cyclic-AMP-responsive-element-binding protein) and ATF family of transcription factors[11]. In resting cells, all three ER-stress sensors are maintained in an inactive state through association with the abundant ER chaperone BiP (immunoglobulin-heavy-chain-binding protein; also known as HSPA5 and GRP78). It has been suggested that in conditions of ER stress, BiP is sequestered through binding to unfolded or misfolded polypeptide chains and/or unassembled multisubunit proteins, thereby leading to the release and, consequently, the activation of the ER-stress sensors[12]. Although this model of BiP sequestration by unfolded proteins is consistent with most experimental evidence, it is probably an oversimplification of the complex interactions between diverse signals that are necessary and/or sufficient to activate the UPR[13] (Fig. 1).

The most immediate response to ER stress, following the release of BiP from PERK, is the homodimerization and *trans*-phosphorylation of PERK, allowing PERK to phosphorylate eIF2α. The phosphorylation of eIF2α inhibits the assembly of the 80S ribosome and, consequently, the synthesis of proteins. This pathway promotes cell survival by preventing the influx of additional nascent polypeptides into an already-saturated ER lumen. Indeed, inhibition of PERK-mediated eIF2α phosphorylation reduces cell survival in conditions of ER stress[14]. However, phosphorylation of eIF2α is required for the translation of certain messenger RNAs that contain regulatory sequences, such as the short open reading frames in the 5′-untranslated region of the mRNA encoding the transcription factor ATF4 (refs 15, 16). ATF4 can induce the expression of UPR target genes, which are involved in amino-acid biosynthesis and transport, the oxidative stress response, and ER-stress-induced apoptosis[17].

In response to ER stress, IRE1α autophosphorylates, thereby activating its RNase activity. It then initiates the removal of a 26-base intron from mRNA encoding X-box-binding protein 1 (XBP1), resulting in a translational frameshift and translation of an XBP1 isoform with potent activity as a transcription factor (referred to here as active XBP1)[5,6].

In parallel, when ATF6 is released from BiP, it translocates to the Golgi apparatus, where it is cleaved by the proteases site-1 protease (S1P) and S2P. This process results in the release of a functional (bZIP-containing) fragment of ATF6 into the cytosol. This fragment then migrates to the nucleus and activates transcription[11,18]. Notably, S1P and S2P also cleave ER-associated sterol-regulatory-element-binding proteins (SREBPs), which are required for cholesterol and fatty-acid biosynthesis[18]. Cleaved ATF6 and active XBP1 isoform function mainly in parallel pathways to induce the transcription of genes encoding ER chaperones and enzymes that promote protein folding, maturation, secretion and ER-associated protein degradation[19,20]. However, if the cell fails to resolve the protein-folding defect and restore homeostasis in the ER, the UPR will initiate apoptosis, to protect the organism by removing the stressed cells that produce misfolded or malfunctioning proteins[5,6]. ER-stress-induced apoptosis is mediated largely by CHOP, a transcription factor that is homologous to C/EBP (CCAAT/enhancer-binding protein) and is downstream of the PERK–eIF2α–ATF4 pathway and the ATF6 pathway in the UPR. Although deletion of the gene encoding CHOP is known to protect cells against ER-stress-induced apoptosis, the mechanism by which CHOP induces apoptosis remains obscure. CHOP has been shown, however, to induce the expression of numerous pro-apoptotic factors (including DR5, TRB3, BIM and GADD34), which promote protein synthesis and oxidative stress in stressed cells[21–24].

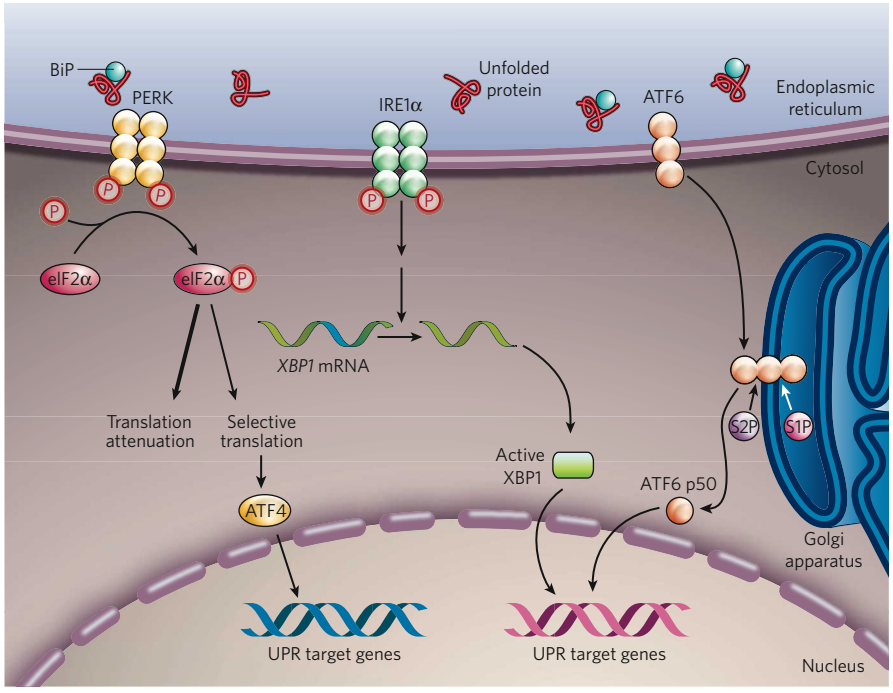


Figure 1 | The mammalian UPR pathways. In non-stressed cells (not shown), the ER chaperone BiP binds to the luminal domains of the ER-stress sensors IRE1α, PERK and ATF6, maintaining these proteins in an inactive state. During ER stress (shown), BiP preferentially binds to unfolded or misfolded proteins, thus driving the equilibrium of BiP binding away from IRE1α, PERK and ATF6. These three proteins are the initiators of the three main signalling cascades of the UPR. The release of BiP results in the activation of PERK, through PERK homodimerization and *trans*-autophosphorylation. Activated PERK then phosphorylates the translation-initiation factor eIF2α, reducing the overall frequency of messenger RNA translation initiation. However, selected mRNAs, such as *ATF4* mRNA, are preferentially translated in the presence of phosphorylated eIF2α. ATF4 activates the transcription of UPR target genes encoding factors involved in amino-acid biosynthesis, the antioxidative-stress response and apoptosis. The release of BiP also allows IRE1α to dimerize, activating its protein-kinase activity (through autophosphorylation) and its endoribonuclease activity. IRE1α then removes a 26-base intron from *XBP1* mRNA. The spliced *XBP1* mRNA encodes a potent transcription factor that translocates to the nucleus, activating the expression of UPR target genes. The release of BiP from ATF6 allows ATF6 to translocate to the Golgi apparatus, where it is cleaved by the proteases S1P and S2P, yielding an active cytosolic ATF6 fragment (ATF6 p50). This fragment migrates to the nucleus, activating the transcription of UPR target genes. S1P, site-1 protease; S2P, site-2 protease; XBP1, X-box-binding protein 1.

Pathways that connect ER stress to inflammation

In addition to the UPR, other signalling pathways radiate from the ER to the mitochondria and nucleus, and possibly to other organelles. A growing body of evidence suggests that the signalling pathways in the UPR and inflammation are interconnected through various mechanisms, including the production of reactive oxygen species (ROS), the release of calcium from the ER, the activation of the transcription factor nuclear factor-κB (NF-κB) and the mitogen-activated protein kinase (MAPK) known as JNK (JUN N-terminal kinase), and the induction of the acute-phase response.

Oxidative protein folding and accumulation of ROS

ROS are small molecules that are highly reactive as a result of the presence of unpaired electrons. ROS are important mediators of inflammation[25], and recent findings have linked ER stress to the generation and accumulation of intracellular ROS, a state commonly referred to as oxidative stress. The folding of proteins into the correct conformations in the ER is an energy-consuming process, and oxidizing conditions are required for the formation of intramolecular and intermolecular disulphide bonds[26]. Electron transport during disulphide-bond formation is driven by a protein relay that involves two ER-resident enzymes: protein disulphide isomerase (PDI) and ER oxidoreductin 1 (ERO1)[27] (Fig. 2). PDI directly accepts electrons, resulting in the oxidation of cysteine residues and the formation of disulphide bonds. ERO1 then uses a flavin-dependent reaction to transfer electrons from PDI to molecular oxygen, thereby oxidizing PDI. Although it provides a robust driving force for disulphide-bond formation, the use of molecular oxygen as the terminal electron recipient leads to the production of ROS[27]. Furthermore, additional oxidative stress can result from the depletion of reduced glutathione, because reduced glutathione is consumed in reactions that reduce unstable and improperly formed disulphide bonds[28]. Therefore, an increase in the protein-folding load in the ER can lead to the accumulation of ROS, which might initiate an inflammatory response.

Importantly, cells have evolved mechanisms to limit the accumulation of ROS in response to ER stress. The PERK pathway of the UPR can activate an antioxidant program by preferentially translating mRNA encoding the bZIP-containing transcription factor ATF4 and by phosphorylating NRF2 (nuclear factor-erythroid-derived 2 (NF-E2)-related factor 2), another bZIP-containing transcription factor[17,29]. After PERK-mediated phosphorylation, NRF2 translocates to the nucleus and activates the transcription of a set of antioxidant and oxidant-detoxifying enzymes, including NAD(P)H–quinone oxidoreductase, haem oxygenase 1 and glutathione *S*-transferase[30,31]. In addition, NRF2 and ATF4 each induce the transcription of genes whose products maintain the cellular level of glutathione, the main redox buffer in the cell[17,29,32]. The overall antioxidant effect of the PERK pathway is supported by the finding that a potent ER-stress-inducing chemical, tunicamycin, induces only weak accumulation of ROS in wild-type cells, whereas this treatment induces a toxic accumulation of ROS in cells that lack PERK[17,32].

ER-associated NF-κB activation and the PERK pathway

NF-κB is a key transcriptional regulator that has a central role in the onset of inflammation[33]. In the absence of inflammatory stimuli, NF-κB remains in an inactive state through binding to a member of the family of inhibitors of NF-κB (IκB), which are constitutively expressed. Activation of NF-κB is initiated by signal-induced phosphorylation of IκB, which is subsequently degraded. The degradation of IκB exposes a nuclear-localization signal in NF-κB, allowing NF-κB to translocate to the nucleus, where it induces the transcription of numerous inflammatory genes. An increase in the ER protein-folding load (for example, during viral infection) has been shown to result in the activation of NF-κB[34,35]. However, the details of the mechanism by which NF-κB is activated in these conditions are poorly understood. Experiments using calcium chelators and antioxidants indicate that, together, these signals contribute to the activation of NF-κB in response to ER stress[36]. Therefore, ER-associated NF-κB activation might result from the oxidative stress of excessive protein folding and/or from an ER-stress-mediated leakage of calcium into the cytosol[37]. In addition, in response to ER stress, the UPR can directly promote NF-κB activation through a PERK–eIF2α-mediated attenuation of translation. Because the half-life of IκB is much shorter than that of NF-κB, attenuating translation increases the ratio of NF-κB to IκB, thereby freeing NF-κB to translocate to the nucleus in response to ER stress[38] (Fig. 3). This effect has been observed in cells treated with reagents that induce ER stress and in cells irradiated with ultraviolet light, both of which activate the PERK pathway of the UPR[38,39].

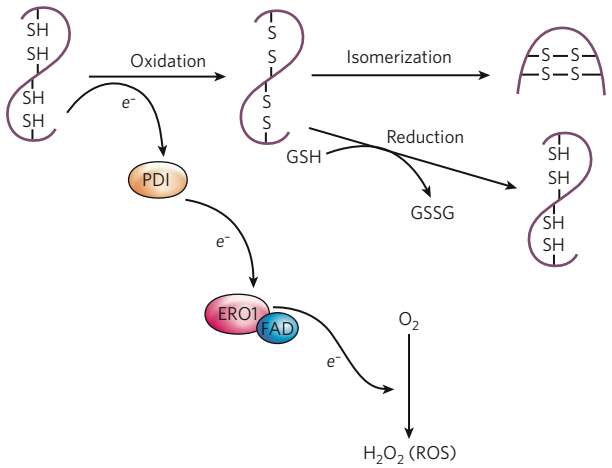


Figure 2 | Oxidative protein folding. The formation of disulphide bonds in proteins in the ER is driven by the enzymes PDI and ERO1. ERO1 operates in association with the flavin FAD, which is synthesized in the cytosol but can readily enter the ER lumen. PDI accepts electrons (e^-) from protein-folding substrates, thereby oxidizing the thiol (SH) groups in the protein's cysteine residues and resulting in the formation of disulphide bonds. ERO1 uses an FAD-dependent reaction to transfer electrons from PDI to molecular oxygen (O_2), resulting in the production of ROS in the form of hydrogen peroxide (H_2O_2). Reduced glutathione (GSH) can assist in disulphide-bond reduction, which occurs when there is a overload of proteins to fold or an accumulation of misfolded proteins, and results in the production of oxidized glutathione (GSSG). In addition, reduced PDI can mediate a reduction of mispaired thiol groups in oxidized protein-folding substrates, functioning as an isomerase. Because the activity of ERO1 is modulated by the amount of FAD in the ER, disulphide-bond formation is linked to the nutritional and/or metabolic status of the cell.

IRE1α-mediated NF-κB and JNK activation

In mammals, IRE1α might be important for integrating ER-stress signalling with inflammatory-response signalling. This is thought to occur in the following manner. In response to ER stress, the autophosphorylation of IRE1α induces a conformational change in its cytosolic domain, which can then bind to the adaptor protein tumour-necrosis factor-α (TNF-α)-receptor-associated factor 2 (TRAF2)[40]. The IRE1α–TRAF2 complex can recruit IκB kinase (IKK), which phosphorylates IκB, leading to the degradation of IκB and the nuclear translocation of NF-κB[41] (Fig. 3). Consistent with these observations, ER-stress-induced NF-κB activation and production of the inflammatory cytokine TNF-α are impaired in mouse embryonic fibroblasts that lack IRE1α[41]. The IRE1α–TRAF2 complex can also recruit the protein kinase JNK, leading to the activation of JNK. Activated JNK induces the expression of inflammatory genes by phosphorylating the transcription factor activator protein 1 (AP1)[42]. Given that JNK activation in response to ER stress is impaired in mouse embryonic fibroblasts that lack IRE1α, IRE1α might provide a link between ER stress and inflammation[40]. Taking these findings together, the formation of the IRE1α–TRAF2 complex seems to be crucial for activating both JNK and NF-κB in response to ER stress. Further studies will be needed to identify how ER-stress-induced signalling involving these two factors, JNK and NF-κB, might be integrated and/or synergize to regulate inflammation, metabolism, cell survival and apoptosis.

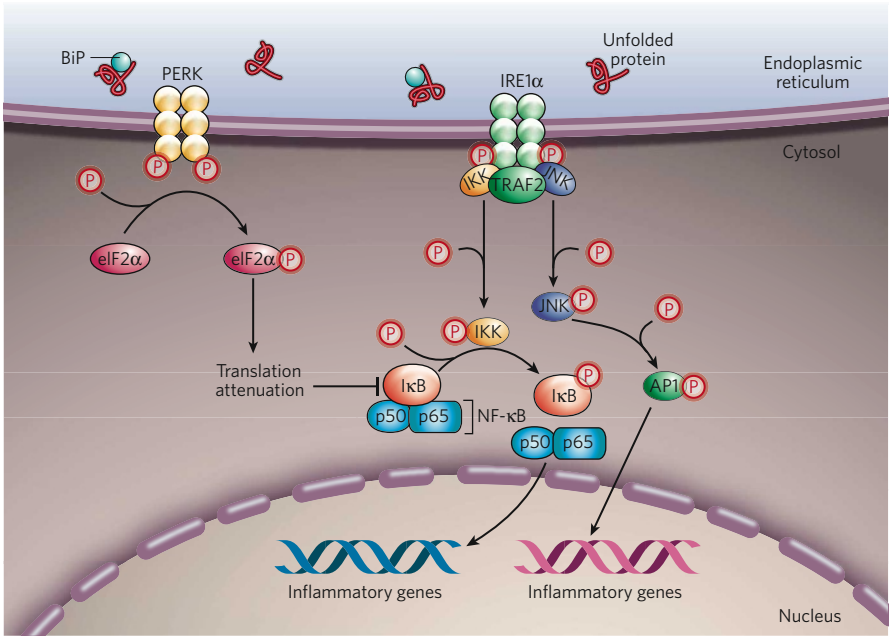


Figure 3 | Proposed models for UPR-mediated JNK and NF-κB activation. In response to ER stress, PERK mediates a general repression of mRNA translation by phosphorylating eIF2α. Because IκB has a shorter half-life than NF-κB, PERK-mediated translational attenuation shifts the ratio of IκB to NF-κB, thereby freeing NF-κB to translocate to the nucleus. In addition, in response to ER stress, the cytoplasmic domain of phosphorylated IRE1α can recruit tumour-necrosis factor-α (TNF-α)-receptor-associated factor 2 (TRAF2). The IRE1α–TRAF2 complex interacts with JNK and/or IκB kinase (IKK), activating these protein kinases. Activated JNK phosphorylates the transcription factor activator protein 1 (AP1). Activated IKK phosphorylates IκB, initiating the degradation of IκB and thereby leading to NF-κB activation. Activated NF-κB and AP1 then migrate to the nucleus, where they induce the transcription of genes involved in the inflammatory response.

The acute-phase response

Regulated intramembrane proteolysis (RIP) is a process by which ER-resident bZIP-containing transcription factors (including SREBPs and ATF6) traffick from the ER to the Golgi apparatus, where they are cleaved, releasing functional isoforms[18,43]. Recently, CREBH, another RIP-regulated bZIP-containing transcription factor, was identified to mediate the acute-phase response in the liver[44]. CREBH is expressed mainly by hepatocytes, and its expression is highly induced by inflammatory cytokines, such as TNF-α, interleukin 1β (IL-1β) and IL-6. When ER stress occurs, CREBH is activated and mediates the acute-phase response in the liver[44] (Fig. 4). CREBH is activated through translocation from the ER to the Golgi apparatus, where it is cleaved by S1P and S2P. An N-terminal fragment of CREBH is released into the cytosol, and this fragment translocates to the nucleus, where it can induce transcription. In the mouse liver, inflammatory cytokines and bacterial lipopolysaccharide (LPS) each induces ER stress and leads to such cleavage of CREBH[44]. However, CREBH does not induce the expression of genes involved in the UPR. Instead, it binds to a DNA-sequence motif in the promoter regions of a subset of acute-phase-response genes, including those encoding serum amyloid P component and C-reactive protein[44]. Further studies are required to elucidate how the trafficking of CREBH from the ER is regulated. In addition, targeted deletion of the gene encoding CREBH should identify the significance of this ER-stress signalling pathway in the inflammatory response.

Factors at the crossroads of inflammation and ER stress

Accumulating evidence suggests that there is extensive cross-talk between the inflammatory response and the ER-stress response. Inflammation can be triggered by a chronic excess of metabolic factors (such as lipids, glucose and cytokines) and/or neurotransmitters. In many physiological or pathological settings, these stimuli can also elicit ER stress, which further disrupts metabolic functions, thereby causing more inflammation. Such vicious cycles could exacerbate inflammatory stress signalling (that is, the signalling pathways that integrate stress and inflammation), as well as metabolic deterioration, in specialized cells such as macrophages, β-cells (in the pancreas) and adipocytes. Moreover, intracellular calcium and free radicals (such as ROS and nitric oxide) are crucial for integrating inflammatory responses, metabolic responses and ER-stress responses, and dynamic signalling by these factors relies on there being functional interactions between the ER and mitochondria.

Calcium and free radicals

The oxidation state and concentration of calcium in the ER lumen crucially affect polypeptide folding, as well as chaperone function. The calcium concentration in the ER is many thousand-fold greater than that in the cytosol[45]. The calcium concentration in the ER is regulated by ATP-dependent uptake of calcium into the ER and receptor-mediated release of calcium from the ER. An accumulation of misfolded proteins in the ER can cause calcium to leak from the ER, possibly through inositol-trisphosphate receptors[37]. The calcium released from the ER is concentrated in the matrix of the mitochondria and causes depolarization of the inner mitochondrial membrane, disrupting electron transport and increasing ROS production[46] (Fig. 5). Mitochondrial ROS can further increase calcium release from the ER by sensitizing ER calcium-release channels and causing protein misfolding. In addition, during oxidative protein folding in the ER, reducing equivalents are transferred from thiol groups in protein-folding substrates to molecular oxygen, thus producing membrane-permeable hydrogen peroxide (an ROS). Through this forward cycle, calcium release, ROS production and protein misfolding function together to activate calcium-dependent protein kinases, as well as JNK and NF-κB, leading to inflammatory responses and even cell death[47].

In addition to ROS, reactive nitrogen species also contribute to inflammation and ER stress. Nitric oxide is a highly reactive, uncharged, membrane-permeable molecule that functions as a signal in many regulatory processes, such as blood-vessel dilation, immune responses and neurotransmission. Nitric oxide can react with superoxide to form peroxynitrite, and with thiols and metal centres in proteins to form nitrosyl adducts[48]. It has also been shown to modify the active site of PDI, thereby interfering with disulphide-bond formation and resulting in the accumulation of misfolded proteins in the ER[49]. In addition, excessive production of nitric oxide can alter the oxidative state and calcium concentration in the ER and disrupt the electron-transport chain, causing ER stress and ROS production[50,51].

Metabolic factors

Several reports indicate that inflammatory cytokines can cause ER stress and therefore activate the UPR. For example, TNF-α causes ER stress, activating PERK, IRE1α and ATF6 in fibrosarcoma cells[52]. In addition, TNF-α, IL-1β and/or IL-6 can induce ER stress in hepatocytes, leading to the activation of CREBH, which then mediates an acute-phase response[44]. And the presence of the T-cell-derived cytokine interferon-γ (IFN-γ) has been associated with PERK activation and

ER-stress-induced apoptosis in oligodendrocytes (cells that produce large amounts of myelin in the nervous system)[53]. Although the mechanism by which cytokines induce ER stress is not completely understood, experimental evidence supports the idea that cytokines trigger the release of calcium from the ER and the accumulation of ROS, which interfere with protein folding and mitochondrial metabolism[52,53].

In addition to cytokines, excessive amounts of metabolic factors, such as cholesterol, non-esterified fatty acids, glucose, homocysteine and neurotransmitters, can also induce both the ER-stress response and the inflammatory response in a variety of cell types[54–58]. There is evidence to support the idea that the presence of large amounts of these metabolic factors can stimulate the release of calcium from the ER, the production of free radicals, and ER stress. But the molecular links between metabolic-factor excess, ER stress and inflammation are not well defined.

The UPR and inflammation in health and disease

The cross-talk between the UPR and inflammation is exemplified in cell types that have metabolic or immune functions. These cell types include hepatocytes, β-cells, adipocytes, macrophages and oligodendrocytes. Because these specialized cell types require the trafficking of large amounts of 'cargo' through the ER, they are extremely sensitive to alterations in metabolism and/or ER homeostasis. Metabolic conditions such as lipid accumulation, increased glucose levels or excessive amounts of cytokines can trigger calcium release from the ER and ROS production in these cells, leading to ER stress and inflammation (Fig. 6). A wealth of evidence from *in vitro* studies suggests that pathological conditions that interfere with ER homeostasis and/or mitochondrial metabolism result in chronic activation of the UPR and in inflammation. Recent observations indicate that the molecular link between ER-stress responses and inflammatory responses might be mediated by activation of two signalling molecules involved in inflammatory responses, JNK and NF-κB. The coupling of the UPR and inflammation in specialized cells and tissues might be fundamental to the pathogenesis of metabolic, neurodegenerative and infectious diseases. In this section, we describe some of the compelling evidence that prolonged activation of the UPR and inflammation are integrated and 'conspire' in the pathogenesis of disease.

Obesity and type 2 diabetes

The ER centrally controls cellular metabolism by regulating protein synthesis and secretion, as well as triglyceride and cholesterol biosynthesis. Metabolic conditions such as insulin resistance and reduced glucose utilization are associated with the development of metabolic syndrome, and these processes are regulated by numerous mechanisms, including the UPR, JNK activation, NF-κB activation and apoptosis[2,59]. Obesity and type 2 diabetes, whether caused by lifestyle factors or genetic deficiency, result in conditions that increase the demand on the ER. This is particularly clear in the liver, adipose tissue and pancreas, where changes in tissue architecture, increases in protein synthesis, and perturbations in cellular energy fluxes occur[2]. Indeed, ER dysfunction has been linked to increased JNK activity, NF-κB activation and insulin resistance[40,41,60,61].

In normal conditions, activated insulin receptors phosphorylate tyrosine residues on proximal signalling molecules, such as insulin-receptor substrate 1 (IRS1), that transmit the effects of insulin by interacting with other cytosolic molecules. Insulin resistance can result from JNK-mediated phosphorylation of serine residues in IRS1, which inhibits the phosphorylation of IRS1 on tyrosine residues[62–64]. In the liver and adipose tissues of obese animals, PERK and IRE1α, and their downstream effectors, have been found to be activated[60]. Because activated IRE1α can recruit TRAF2 and trigger JNK activation when ER stress occurs[40], it has been proposed that IRE1α links ER stress and JNK-mediated serine phosphorylation of IRS1, causing peripheral insulin resistance. Consistent with this hypothesis, when mice that lacked one allele of *Xbp1* (*Xbp1*$^{+/-}$ mice) were fed a high-fat diet, the liver and adipose tissues showed increased activation of PERK, IRE1α and JNK, and dysregulated phosphorylation of IRS1, coupled with insulin resistance[60]. It is possible that reduced signalling through XBP1 compromises protein folding and thereby causes insulin resistance, although further studies are required to validate this hypothesis. In addition, because both IRE1α and PERK activation can lead to NF-κB activation (through IKK activation and translation attenuation, respectively; Fig. 3), further studies are required to elucidate the significance of IRE1α and PERK in coordinating the activation of JNK and NF-κB, as well as the impact of this coordinated activation on the insulin resistance and inflammation that are associated with obesity and type 2 diabetes.

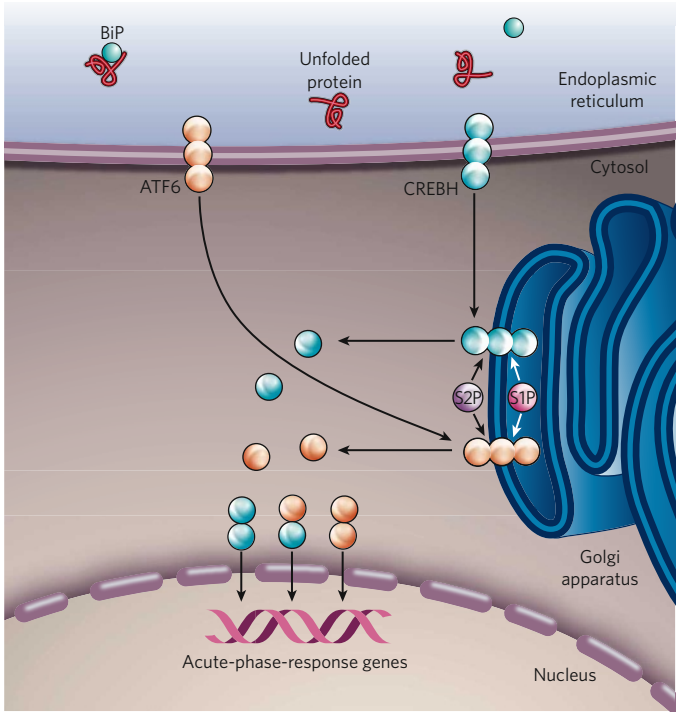


Figure 4 | The ER-stress-induced acute-phase response. When inflammatory cytokines, such as TNF-α, IL-1β and IL-6, are present in the extracellular environment, the gene encoding CREBH is transcribed (not shown). CREBH, similar to ATF6, is a bZIP-containing transcription factor that is localized to the ER membrane. CREBH, however, is mainly expressed by hepatocytes, whereas ATF6 is expressed by all cell types. In conditions of ER stress, such as those caused by inflammatory cytokines or the bacterial component lipopolysaccharide (LPS), CREBH translocates to the Golgi apparatus, where it is cleaved by the proteases S1P and S2P, releasing a cytosolic fragment. ER stress also activates the UPR sensor ATF6 by regulated intramembrane proteolysis. Activated CREBH and ATF6 can then form homodimers or heterodimers and migrate to the nucleus, where they activate the transcription of the genes encoding serum amyloid P component and C-reactive protein, which mediate the acute-phase response.

Atherosclerosis

Atherosclerosis, the leading cause of cardiovascular disease, is an inflammatory disease in which immune mechanisms interact with metabolic risk factors, causing lesions to develop in the arterial vasculature. Cholesterol deposition by macrophages, inflammation and cell death are crucial contributors to the formation and progression of these lesions, resulting in the acute occlusion of blood vessels[65]. Recent evidence suggests that the UPR and inflammation underlie the development of atherosclerotic lesions.

The accumulation of free cholesterol in the ER membranes of macrophages causes calcium release, UPR activation and CHOP-induced apoptosis[54]. This loading of macrophages with free cholesterol activates NF-κB and the MAPKs p38, extracellular-signal-regulated kinase 1 (ERK1) and ERK2, and JNK, thereby inducing the expression of genes encoding inflammatory cytokines (including TNF-α and IL-6)[66]. In these conditions, JNK and NF-κB activation might be mediated, in part, through PERK and IRE1α[66]. Interestingly, CHOP, which is mainly produced by way of the PERK pathway of the UPR, is required for IL-6

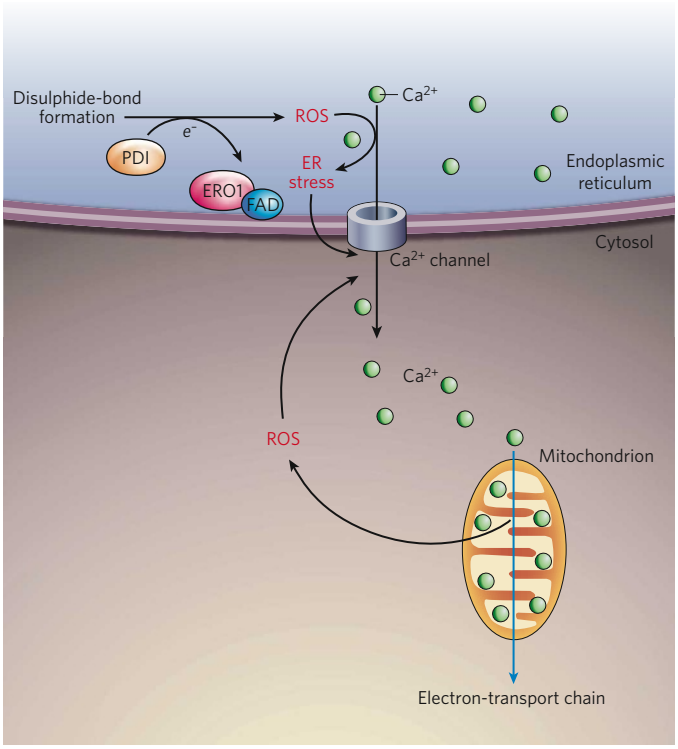


Figure 5 | The role of calcium and ROS in the UPR and inflammation. Protein folding is an oxidative process that generates ROS. ROS can target chaperones (not shown) and ER-based calcium (Ca^{2+}) channels, leading to the release of calcium from the ER into the cytosol and ER-stress signalling. Calcium released from the ER is concentrated in the inner matrix of the mitochondria, where it disrupts the electron-transport chain, thereby leading to the production of more ROS. These mitochondrially produced ROS can further exacerbate calcium release from the ER, resulting in the accumulation of ROS to a toxic level. Furthermore, perturbation of ER calcium homeostasis can disrupt the protein-folding process in the ER, which, in turn, causes ER stress, induces the UPR and generates more ROS.

production and for full activation of ERK1 and ERK2 in response to loading with free cholesterol. The connections between the UPR, ER-stress-induced apoptosis and inflammation might help to explain the link between free-cholesterol accumulation and inflammation in the vulnerability of lesions to rupture in advanced atherosclerosis. In addition to the free-cholesterol loading of macrophages, oxidized lipids (such as oxidized low-density lipoprotein and its bioreactive component, oxidized 1-palmitoyl-2-arachidonyl-sn-3-glycero-phosphorylcholine) can result in ER stress and UPR activation in human aortic endothelial cells[67]. The UPR is also activated in human atherosclerotic lesions, where oxidized phospholipids have accumulated. Furthermore, *in vitro* studies have shown that the ER-stress-induced transcription factors ATF4 and XBP1 are required for the production of the inflammatory cytokine IL-6 and the chemokines IL-8 (also known as CXC-chemokine ligand 8 (CXCL8)), CC-chemokine ligand 2 (CCL2) and CXCL3 by human aortic endothelial cells in the basal state and on accumulation of oxidized lipids[67]. Together, these studies suggest that UPR signalling is an important mediator of vascular inflammation and possibly of the endothelial-cell dysfunction that is observed in atherosclerosis.

Neurodegenerative diseases

Most acute and chronic neurodegenerative diseases involve inflammation, although the source of the inflammatory response is poorly characterized[68]. These diseases, including Alzheimer's disease, Parkinson's disease, multiple sclerosis and diseases that result from the expansion of a polyglutamine repeat, are associated with protein aggregation and are characterized by abnormal neuronal physiology and neural-cell death[69]. Recent studies suggest that the protein aggregates associated with these diseases might inhibit the proteasome, thereby preventing ER-associated protein degradation (ERAD) and leading to the accumulation of unfolded proteins in the ER[70,71]. However, there is limited evidence from studies of animal models or humans to support the idea that the pathology associated with these diseases results from defects in ERAD that cause ER stress. Intriguingly, mutations in genes that have functions linked to ERAD and/or mitochondrial function can cause Parkinson's disease in humans[72]. In addition, deletion of the gene encoding the pro-apoptotic UPR-induced transcription factor CHOP was reported to protect against apoptosis in a neurotoxin-induced mouse model of Parkinson's disease[73]. However, in a mouse model, brain-specific deletion of the gene encoding XBP1, a transcriptional activator of genes whose products are involved in ERAD, did not affect the development of prion disease (a family of neurodegenerative diseases caused by prion-protein misfolding)[74]. Although the mechanisms underlying neurodegenerative diseases are still under investigation, it is clear that alterations in protein folding, calcium signals, redox homeostasis and inflammation are prominent features[69,75,76].

Multiple sclerosis is a neurodegenerative disease that is marked by demyelination, oligodendrocyte loss and T-cell activation associated with IFN-γ production[53,77]. In animal models of multiple sclerosis, treatment with IFN-γ was found to induce ER stress in actively myelinating oligodendrocytes, leading to apoptosis of the oligodendrocytes and abnormalities in neuron myelination[78]. By contrast, treatment with IFN-γ has been shown to activate the PERK pathway of the UPR, protecting mature oligodendrocytes against immune-mediated damage[79]. It has been proposed that these divergent responses to IFN-γ depend on the rate of protein synthesis by the oligodendrocytes[79]. In an oligodendrocyte that is actively producing myelin, the increase in protein production stimulated by IFN-γ could convert an adaptive (able to be adjusted), moderate level of ER stress to a destructive, apoptosis-inducing, ER-stress response. By contrast, mature oligodendrocytes produce less protein, so an increase in protein production might not result in such a destructive ER-stress response[79,80]. Thus, regulating the balance between the rates of protein production and the inflammatory stress responses (which integrate inflammation and ER-stress signalling) in oligodendrocytes might be a crucial factor in the development of demyelinating diseases.

Therapeutic potential and future directions

Considerable progress has now been made towards understanding the signalling pathways that integrate the UPR and inflammation and the physiological significance of this connection. Recently, researchers have focused on designing effective therapeutics for inflammatory diseases by modulating the UPR and inflammatory response. However, manipulating the interface between these fundamental biological responses for therapeutic purposes, without causing severe side effects, is a formidable challenge. Because many mediators of cellular stress and inflammation are regulated simultaneously, it is unlikely that a single response that integrates inflammation and ER-stress signalling is responsible for the pathogenesis of a particular disease. Given this complexity, an effective approach would be to seek to re-establish functional homeostasis by modifying integrated biological outcomes rather than targeting single pathways.

Recent studies suggest that preserving or restoring ER function might be therapeutic. Small molecules that are classified as chemical chaperones can facilitate protein folding and protect against ER stress, thus relieving disease symptoms in animal models. For example, in insulin-resistant obese mice, the chemical chaperones 4-phenylbutyric acid and taurine-conjugated ursodeoxycholic acid were found both to reduce the phosphorylation of PERK and IRE1α significantly and to improve glucose tolerance and insulin sensitivity[61]. In addition, another chemical chaperone, the resveratrol tetramer vaticanol B, has been shown to inhibit both the UPR and the inflammatory response by reducing the protein-folding load and maintaining ER-membrane integrity, preventing ER-stress-induced apoptosis[81]. In addition to chemical chaperones, salubrinal, a phosphatase inhibitor, might have therapeutic benefit. Salubrinal can

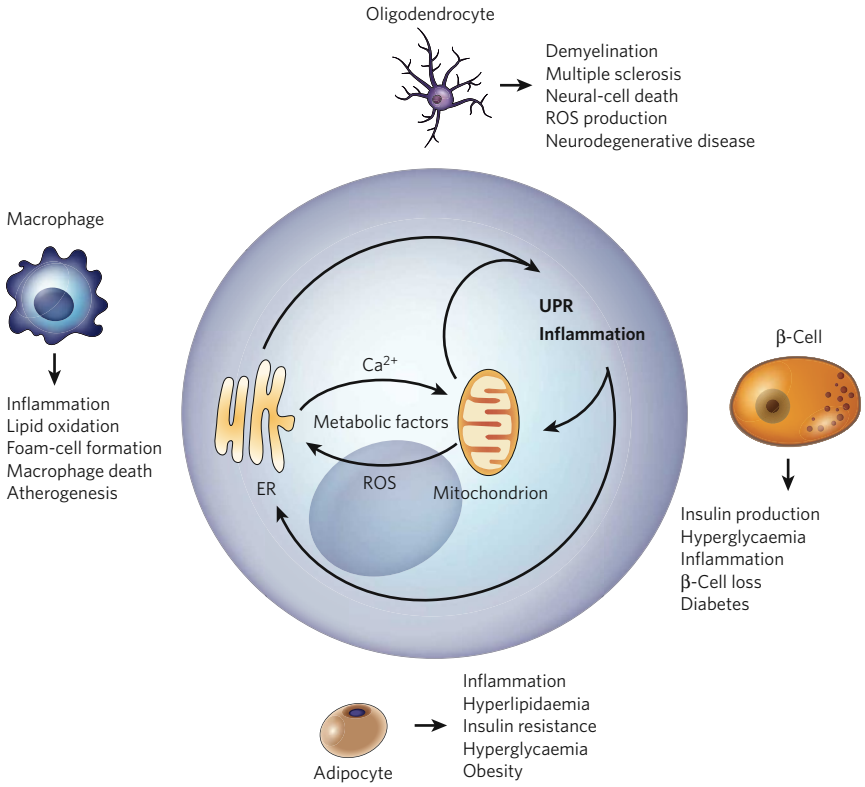


Figure 6 | The 'ER-stress–inflammation' loop in specialized cells. In specialized cells that secrete large amounts of protein — such as macrophages, adipocytes, β-cells and oligodendrocytes — the UPR and inflammatory-response signalling can be triggered by a chronic excess of extracellular and/or intracellular metabolic factors, such as lipids, glucose, cytokines, hormones, non-esterified fatty acids and neurotransmitters. More specifically, such metabolic factors stimulate protein synthesis, calcium signalling and ROS production by targeting the mitochondria and the ER in these cells (Fig. 5). The increased protein-folding demand and the signalling involving calcium and ROS induce the UPR and inflammatory-response signalling, leading to the transcription of genes whose products mount a broader inflammatory response. An excess of metabolic factors can further boost the UPR and inflammation, contributing to impaired lipid and glucose metabolism, insulin resistance and apoptosis. This forward ER-stress–inflammation loop could also further promote inflammatory stress signalling and contribute to the metabolic deterioration that is associated with atherosclerosis, obesity, type 2 diabetes and neurodegenerative diseases, depending on the cell type involved.

protect cells against ER-stress-induced apoptosis by selectively inhibiting the dephosphorylation of eIF2α such that further protein synthesis and accumulation in the ER is inhibited[82].

Future studies will need to address the many open questions about the physiological significance of the various ER-stress signalling pathways in mediating inflammatory responses. The knowledge gained by such studies will improve the overall understanding of how inflammatory diseases develop and indicate how they might be treated with pharmacological interventions that modulate ER stress and inflammation. ■

1. Charo, I. F. & Ransohoff, R. M. The many roles of chemokines and chemokine receptors in inflammation. *N. Engl. J. Med.* **354,** 610–621 (2006).
2. Hotamisligil, G. S. Inflammation and metabolic disorders. *Nature* **444,** 860–867 (2006).
3. Hansson, G. K. & Libby, P. The immune response in atherosclerosis: a double-edged sword. *Nature Rev. Immunol.* **6,** 508–519 (2006).
4. Kaufman, R. J. Stress signaling from the lumen of the endoplasmic reticulum: coordination of gene transcriptional and translational controls. *Genes Dev.* **13,** 1211–1233 (1999).
5. Ron, D. & Walter, P. Signal integration in the endoplasmic reticulum unfolded protein response. *Nature Rev. Mol. Cell Biol.* **8,** 519–529 (2007).
6. Schroder, M. & Kaufman, R. J. The mammalian unfolded protein response. *Annu. Rev. Biochem.* **74,** 739–789 (2005).
7. Mori, K., Ma, W., Gething, M. J. & Sambrook, J. A transmembrane protein with a cdc2^{+}/CDC28-related kinase activity is required for signaling from the ER to the nucleus. *Cell* **74,** 743–756 (1993).
8. Cox, J. S., Shamu, C. E. & Walter, P. Transcriptional induction of genes encoding endoplasmic reticulum resident proteins requires a transmembrane protein kinase. *Cell* **73,** 1197–1206 (1993).
9. Shi, Y. *et al.* Identification and characterization of pancreatic eukaryotic initiation factor 2 α-subunit kinase, PEK, involved in translational control. *Mol. Cell. Biol.* **18,** 7499–7509 (1998).
10. Harding, H. P., Zhang, Y. & Ron, D. Protein translation and folding are coupled by an endoplasmic-reticulum-resident kinase. *Nature* **397,** 271–274 (1999).
11. Haze, K., Yoshida, H., Yanagi, H., Yura, T. & Mori, K. Mammalian transcription factor ATF6 is synthesized as a transmembrane protein and activated by proteolysis in response to endoplasmic reticulum stress. *Mol. Biol. Cell* **10,** 3787–3799 (1999).
12. Bertolotti, A., Zhang, Y., Hendershot, L. M., Harding, H. P. & Ron, D. Dynamic interaction of BiP and ER stress transducers in the unfolded-protein response. *Nature Cell Biol.* **2,** 326–332 (2000).
13. Kohno, K. How transmembrane proteins sense endoplasmic reticulum stress. *Antioxid. Redox Signal.* **9,** 2295–2303 (2007).
14. Harding, H. P., Zhang, Y., Bertolotti, A., Zeng, H. & Ron, D. Perk is essential for translational regulation and cell survival during the unfolded protein response. *Mol. Cell* **5,** 897–904 (2000).
15. Lu, P. D., Harding, H. P. & Ron, D. Translation reinitiation at alternative open reading frames regulates gene expression in an integrated stress response. *J. Cell Biol.* **167,** 27–33 (2004).
16. Yaman, I. *et al.* The zipper model of translational control: a small upstream ORF is the switch that controls structural remodeling of an mRNA leader. *Cell* **113,** 519–531 (2003).
17. Harding, H. P. *et al.* An integrated stress response regulates amino acid metabolism and resistance to oxidative stress. *Mol. Cell* **11,** 619–633 (2003).
18. Ye, J. *et al.* ER stress induces cleavage of membrane-bound ATF6 by the same proteases that process SREBPs. *Mol. Cell* **6,** 1355–1364 (2000).
19. Yamamoto, K. *et al.* Transcriptional induction of mammalian ER quality control proteins is mediated by single or combined action of ATF6α and XBP1. *Dev. Cell* **13,** 365–376 (2007).
20. Wu, J. *et al.* ATF6α optimizes long-term endoplasmic reticulum function to protect cells from chronic stress. *Dev. Cell* **13,** 351–364 (2007).
21. Ohoka, N., Yoshii, S., Hattori, T., Onozaki, K. & Hayashi, H. TRB3, a novel ER stress-inducible gene, is induced via ATF4–CHOP pathway and is involved in cell death. *EMBO J.* **24,** 1243–1255 (2005).
22. Yamaguchi, H. & Wang, H. G. CHOP is involved in endoplasmic reticulum stress-induced apoptosis by enhancing DR5 expression in human carcinoma cells. *J. Biol. Chem.* **279,** 45495–45502 (2004).
23. Puthalakath, H. *et al.* ER stress triggers apoptosis by activating BH3-only protein Bim. *Cell* **129,** 1337–1349 (2007).
24. Song, B., Scheuner, D., Ron, D., Pennathur, S. & Kaufman, R. Genetic deletion of C/EBP homologous protein CHOP reduces oxidative stress, improves β cell function, and prevents diabetes. *J. Clin. Invest.* (in the press).
This report describes how the ER-stress-induced pro-apoptotic factor CHOP is involved in oxidative stress and β-cell death.
25. Raha, S. & Robinson, B. H. Mitochondria, oxygen free radicals, disease and ageing. *Trends Biochem. Sci.* **25,** 502–508 (2000).
26. Tu, B. P. & Weissman, J. S. Oxidative protein folding in eukaryotes: mechanisms and consequences. *J. Cell Biol* .**164,** 341–346 (2004).
27. Tu, B. P. & Weissman, J. S. The FAD- and O_2-dependent reaction cycle of Ero1-mediated oxidative protein folding in the endoplasmic reticulum. *Mol. Cell* **10,** 983–994 (2002).
28. Cuozzo, J. W. & Kaiser, C. A. Competition between glutathione and protein thiols for disulphide-bond formation. *Nature Cell Biol.* **1,** 130–135 (1999).
References 27 and 28 provide insights into how protein folding in the ER leads to the production of ROS.
29. Cullinan, S. B. *et al.* Nrf2 is a direct PERK substrate and effector of PERK-dependent cell survival. *Mol. Cell. Biol.* **23,** 7198–7209 (2003).
30. Mathers, J. *et al.* Antioxidant and cytoprotective responses to redox stress. *Biochem. Soc. Symp.* **71,** 157–176 (2004).
31. Zhang, D. D. Mechanistic studies of the Nrf2-Keap1 signaling pathway. *Drug Metab. Rev.* **38,** 769–789 (2006).
32. Cullinan, S. B. & Diehl, J. A. PERK-dependent activation of Nrf2 contributes to redox homeostasis and cell survival following endoplasmic reticulum stress. *J. Biol. Chem.* **279,** 20108–20117 (2004).
33. Rius, J. *et al.* NF-κB links innate immunity to the hypoxic response through transcriptional regulation of HIF-1α. *Nature* **453,** 807–811 (2008).
34. Pahl, H. L. & Baeuerle, P. A. Expression of influenza virus hemagglutinin activates transcription factor NF-κB. *J. Virol.* **69,** 1480–1484 (1995).
35. Meyer, M. *et al.* Hepatitis B virus transactivator MHBst: activation of NF-κB, selective inhibition by antioxidants and integral membrane localization. *EMBO J.* **11,** 2991–3001 (1992).

36. Pahl, H. L. & Baeuerle, P. A. Activation of NF-κB by ER stress requires both Ca^{2+} and reactive oxygen intermediates as messengers. *FEBS Lett.* **392,** 129–136 (1996).
37. Deniaud, A. *et al.* Endoplasmic reticulum stress induces calcium-dependent permeability transition, mitochondrial outer membrane permeabilization and apoptosis. *Oncogene* **27,** 285–299 (2008).
This paper shows that protein misfolding in the ER causes calcium to leak into the cytosol, resulting in the outer membrane of mitochondria becoming more permeable.
38. Deng, J. *et al.* Translational repression mediates activation of nuclear factor κB by phosphorylated translation initiation factor 2. *Mol. Cell. Biol.* **24,** 10161–10168 (2004).
39. Wu, S. *et al.* Ultraviolet light activates NFκB through translational inhibition of IκBα synthesis. *J. Biol. Chem.* **279,** 34898–34902 (2004).
References 38 and 39 show that NF-κB is activated by the PERK pathway of the UPR.
40. Urano, F. *et al.* Coupling of stress in the ER to activation of JNK protein kinases by transmembrane protein kinase IRE1. *Science* **287,** 664–666 (2000).
This paper shows how ER stress activates JNK by way of IRE1α.
41. Hu, P., Han, Z., Couvillon, A. D., Kaufman, R. J. & Exton, J. H. Autocrine tumor necrosis factor α links endoplasmic reticulum stress to the membrane death receptor pathway through IRE1α-mediated NF-κB activation and down-regulation of TRAF2 expression. *Mol. Cell. Biol.* **26,** 3071–3084 (2006).
42. Davis, R. J. Signal transduction by the JNK group of MAP kinases. *Cell* **103,** 239–252 (2000).
43. Brown, M. S., Ye, J., Rawson, R. B. & Goldstein, J. L. Regulated intramembrane proteolysis: a control mechanism conserved from bacteria to humans. *Cell* **100,** 391–398 (2000).
44. Zhang, K. *et al.* Endoplasmic reticulum stress activates cleavage of CREBH to induce a systemic inflammatory response. *Cell* **124,** 587–599 (2006).
This study identifies CREBH, an ER-stress-inducible transcription factor that can mediate the acute-phase response.
45. Berridge, M. J., Bootman, M. D. & Roderick, H. L. Calcium signalling: dynamics, homeostasis and remodelling. *Nature Rev. Mol. Cell Biol.* **4,** 517–529 (2003).
46. Gorlach, A., Klappa, P. & Kietzmann, T. The endoplasmic reticulum: folding, calcium homeostasis, signaling, and redox control. *Antioxid. Redox Signal.* **8,** 1391–1418 (2006).
47. Malhotra, J. D. & Kaufman, R. J. Endoplasmic reticulum stress and oxidative stress: a vicious cycle or a double-edged sword? *Antioxid. Redox Signal.* **9,** 2277–2293 (2007).
48. Stamler, J. S., Singel, D. J. & Loscalzo, J. Biochemistry of nitric oxide and its redox-activated forms. *Science* **258,** 1898–1902 (1992).
49. Uehara, T. *et al.* S-nitrosylated protein-disulphide isomerase links protein misfolding to neurodegeneration. *Nature* **441,** 513–517 (2006).
50. Xu, K. Y., Huso, D. L., Dawson, T. M., Bredt, D. S. & Becker, L. C. Nitric oxide synthase in cardiac sarcoplasmic reticulum. *Proc. Natl Acad. Sci. USA* **96,** 657–662 (1999).
51. Xu, W., Liu, L., Charles, I. G. & Moncada, S. Nitric oxide induces coupling of mitochondrial signalling with the endoplasmic reticulum stress response. *Nature Cell Biol.* **6,** 1129–1134 (2004).
52. Xue, X. *et al.* Tumor necrosis factor α (TNFα) induces the unfolded protein response (UPR) in a reactive oxygen species (ROS)-dependent fashion, and the UPR counteracts ROS accumulation by TNFα. *J. Biol. Chem.* **280,** 33917–33925 (2005).
53. Lin, W., Harding, H. P., Ron, D. & Popko, B. Endoplasmic reticulum stress modulates the response of myelinating oligodendrocytes to the immune cytokine interferon-γ. *J. Cell Biol.* **169,** 603–612 (2005).
54. Feng, B. *et al.* The endoplasmic reticulum is the site of cholesterol-induced cytotoxicity in macrophages. *Nature Cell Biol.* **5,** 781–792 (2003).
55. Maedler, K. *et al.* Glucose-induced β cell production of IL-1β contributes to glucotoxicity in human pancreatic islets. *J. Clin. Invest.* **110,** 851–860 (2002).
56. Kharroubi, I. *et al.* Free fatty acids and cytokines induce pancreatic β-cell apoptosis by different mechanisms: role of nuclear factor-κB and endoplasmic reticulum stress. *Endocrinology* **145,** 5087–5096 (2004).
57. Zhou, J. *et al.* Association of multiple cellular stress pathways with accelerated atherosclerosis in hyperhomocysteinemic apolipoprotein E-deficient mice. *Circulation* **110,** 207–213 (2004).
58. Yamamuro, A., Yoshioka, Y., Ogita, K. & Maeda, S. Involvement of endoplasmic reticulum stress on the cell death induced by 6-hydroxydopamine in human neuroblastoma SH-SY5Y cells. *Neurochem. Res.* **31,** 657–664 (2006).
59. Kaufman, R. J. Orchestrating the unfolded protein response in health and disease. *J. Clin. Invest.* **110,** 1389–1398 (2002).
60. Ozcan, U. *et al.* Endoplasmic reticulum stress links obesity, insulin action, and type 2 diabetes. *Science* **306,** 457–461 (2004).
61. Ozcan, U. *et al.* Chemical chaperones reduce ER stress and restore glucose homeostasis in a mouse model of type 2 diabetes. *Science* **313,** 1137–1140 (2006).
This paper shows that decreasing ER stress improves insulin sensitivity in mice with type 2 diabetes.
62. Hirosumi, J. *et al.* A central role for JNK in obesity and insulin resistance. *Nature* **420,** 333–336 (2002).
63. Aguirre, V. *et al.* Phosphorylation of Ser307 in insulin receptor substrate-1 blocks interactions with the insulin receptor and inhibits insulin action. *J. Biol. Chem.* **277,** 1531–1537 (2002).
64. Tuncman, G. *et al.* Functional *in vivo* interactions between JNK1 and JNK2 isoforms in obesity and insulin resistance. *Proc. Natl Acad. Sci. USA* **103,** 10741–10746 (2006).
65. Williams, K. J. & Tabas, I. Atherosclerosis and inflammation. *Science* **297,** 521–522 (2002).
66. Li, Y. *et al.* Free cholesterol-loaded macrophages are an abundant source of tumor necrosis factor-α and interleukin-6: model of NF-κB- and MAP kinase-dependent inflammation in advanced atherosclerosis. *J. Biol. Chem.* **280,** 21763–21772 (2005).
This paper describes how ER-stress signalling and inflammatory-response signalling are integrated in cholesterol-loaded macrophages.
67. Gargalovic, P. S. *et al.* The unfolded protein response is an important regulator of inflammatory genes in endothelial cells. *Arterioscler. Thromb. Vasc. Biol.* **26,** 2490–2496 (2006).
68. Tansey, M. G., McCoy, M. K. & Frank-Cannon, T. C. Neuroinflammatory mechanisms in Parkinson's disease: potential environmental triggers, pathways, and targets for early therapeutic intervention. *Exp. Neurol.* **208,** 1–25 (2007).
69. Lindholm, D., Wootz, H. & Korhonen, L. ER stress and neurodegenerative diseases. *Cell Death Differ.* **13,** 385–392 (2006).
70. Bence, N. F., Sampat, R. M. & Kopito, R. R. Impairment of the ubiquitin-proteasome system by protein aggregation. *Science* **292,** 1552–1555 (2001).
71. Nishitoh, H. *et al.* ALS-linked mutant SOD1 induces ER stress- and ASK1-dependent motor neuron death by targeting Derlin-1. *Genes Dev.* **22,** 1451–1464 (2008).
72. Wang, H. Q. & Takahashi, R. Expanding insights on the involvement of endoplasmic reticulum stress in Parkinson's disease. *Antioxid. Redox Signal.* **9,** 553–561 (2007).
73. Silva, R. M. *et al.* CHOP/GADD153 is a mediator of apoptotic death in substantia nigra dopamine neurons in an *in vivo* neurotoxin model of parkinsonism. *J. Neurochem.* **95,** 974–986 (2005).
74. Hetz, C. *et al.* Unfolded protein response transcription factor XBP-1 does not influence prion replication or pathogenesis. *Proc. Natl Acad. Sci. USA* **105,** 757–762 (2008).
75. Paschen, W., Aufenberg, C., Hotop, S. & Mengesdorf, T. Transient cerebral ischemia activates processing of *xbp1* messenger RNA indicative of endoplasmic reticulum stress. *J. Cereb. Blood Flow Metab.* **23,** 449–461 (2003).
76. DeLegge, M. H. & Smoke, A. Neurodegeneration and inflammation. *Nutr. Clin. Pract.* **23,** 35–41 (2008).
77. Frohman, E. M., Racke, M. K. & Raine, C. S. Multiple sclerosis — the plaque and its pathogenesis. *N. Engl. J. Med.* **354,** 942–955 (2006).
78. Lin, W. *et al.* Interferon-γ inhibits central nervous system remyelination through a process modulated by endoplasmic reticulum stress. *Brain* **129,** 1306–1318 (2006).
79. Lin, W. *et al.* The integrated stress response prevents demyelination by protecting oligodendrocytes against immune-mediated damage. *J. Clin. Invest.* **117,** 448–456 (2007).
This paper shows that IFN-γ can have a detrimental role or a protective role, mediated by the UPR, depending on the stage of multiple sclerosis.
80. Lees, J. R. & Cross, A. H. A little stress is good: IFN-γ, demyelination, and multiple sclerosis. *J. Clin. Invest.* **117,** 297–299 (2007).
81. Tabata, Y. *et al.* Vaticanol B, a resveratrol tetramer, regulates endoplasmic reticulum stress and inflammation. *Am. J. Physiol. Cell. Physiol.* **293,** C411–C418 (2007).
82. Boyce, M. *et al.* A selective inhibitor of eIF2α dephosphorylation protects cells from ER stress. *Science* **307,** 935–939 (2005).

Acknowledgements We thank J. Mitchell for her efforts in preparing the manuscript. We apologize to those whose work could not be cited because of space limitations. K.Z. is supported by a grant from the American Heart Association (0635423Z). R.J.K. is supported by grants from the National Institutes of Health (DK042394, HL052173 and HL057346) and is an investigator of the Howard Hughes Medical Institute.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence should be addressed to R.J.K. (kaufmanr@umich.edu).

The role of exercise and PGC1α in inflammation and chronic disease

Christoph Handschin[1] & Bruce M. Spiegelman[2]

Inadequate physical activity is linked to many chronic diseases. But the mechanisms that tie muscle activity to health are unclear. The transcriptional coactivator PGC1α has recently been shown to regulate several exercise-associated aspects of muscle function. We propose that this protein controls muscle plasticity, suppresses a broad inflammatory response and mediates the beneficial effects of exercise.

Over the past century, people in the developed world have become less physically active, as a result of the changing nature of work and the availability of machines to replace muscle power. These changes have led to a marked increase in the incidence of many chronic diseases. It is clear that inadequate physical activity can result in obesity, cardiovascular diseases and type 2 diabetes. In addition, a lack of sufficient exercise has also been linked to immune-system dysfunction, pulmonary diseases, musculoskeletal disorders and certain types of cancer and neurological disorder[1] (Box 1). A sedentary lifestyle is therefore a major risk factor for many chronic pathologies.

Inactivity has also been shown unequivocally to increase the morbidity and mortality rates of associated chronic disorders[2,3]. Accordingly, the capacity to exercise is a strong predictor of overall mortality rates, regardless of health status and race[4]. More than 50% of adults in the United States do not exercise enough to benefit their health, and 25% do not engage in any form of physical activity in their leisure time (ref. 5 and see http://www.cdc.gov/nccdphp/dnpa/physical). The trend towards physical inactivity among young people in the Western world is of particular concern[6], and the devastating effects of insufficient physical activity are also observed in the elderly[7]. A decrease in muscle function in the elderly is not only directly linked to sarcopenia (a reduction in skeletal muscle mass owing to ageing) and to the prevalence of several chronic diseases, but also contributes enormously to overall quality of life, by diminishing strength, the ability to carry out daily tasks and social interactions, mobility, cognitive performance and life expectancy[7]. Even in the early elderly years, changes in physical activity have marked consequences for health status and life expectancy. For example, in one study, men of 70 years of age or older who engaged in exercise increased the probability that they would live until age 90 from 44% (for inactive men) to 54% (ref. 8).

Increasing the amount of physical activity effectively prevents the development of many chronic diseases. Moreover, exercise is an excellent therapeutic intervention for conditions such as obesity, type 2 diabetes, neurodegeneration, osteoporosis and sarcopenia[1]. In terms of efficacy, exercise can be as beneficial as the drugs that are prescribed for many of these conditions (for example, for type 2 diabetes)[9].

The mechanisms that mediate the therapeutic effects of exercise and the pathological changes elicited by a sedentary lifestyle remain enigmatic. Here we propose a hypothesis for how inactivity, inflammation and chronic disease are linked at the molecular level, focusing on observations from human studies (except where noted).

Effects of inactivity on inflammation and chronic disease

Many chronic diseases are associated with sterile, persistent, low-grade inflammation (Fig. 1). For example, the infiltration of immune cells into white adipose tissue, and therefore inflammation in this tissue, is closely correlated with the development of insulin resistance and type 2 diabetes[10]. Similarly, activated immune cells and inflammation have an important role in cardiovascular diseases, particularly in the aetiology of atherosclerosis[11,12]. In addition, a systemic increase in the concentration of inflammatory cytokines stimulates the initiation of tumours, as well as their promotion and progression[13,14].

Several neurodegenerative diseases are also linked to inflammation. These diseases are associated with a local inflammatory response in the brain, known as neuroinflammation. Neuroinflammation, for example, affects the activation of glial cells and the subsequent release of inflammatory cytokines such as tumour-necrosis factor-α (TNF-α). These cytokines are thought to promote the death of dopamine-containing neurons in the substantia nigra region of the brain, thereby contributing to the pathology of Parkinson's disease[15,16]. Similarly, interleukin 1β (IL-1β), TNF-α-related apoptosis-inducing ligand (TRAIL) and other cytokines have been postulated to be involved in the aetiology of Alzheimer's disease[17], as has amyloid-β, which itself has inflammatory effects[18]. It should be noted that in addition to the local inflammation found in many neurodegenerative diseases, systemic inflammation further exacerbates these diseases and promotes the progression of neurodegeneration[19].

Physical activity is tightly linked to inflammation and immunity in a complex manner[20]. Regular, moderate exercise reduces the level of systemic inflammation[21]. The mediators of this beneficial effect are unclear, but at least two distinct candidate mechanisms have been identified. First, exercise increases the release of adrenaline, cortisol, growth hormone, prolactin and other factors that have immunomodulatory effects[22]. Second, exercise results in decreased expression of Toll-like receptors at the surface of monocytes, which have been suggested to be involved in mediating systemic inflammation[23]. In contrast to the reduction of chronic inflammation that is afforded by regular, moderate exercise, high-intensity training for a prolonged period results in an increase in systemic inflammation and an increased risk of infection[21]. After such exercise, athletes undergo a transient exercise-induced immunodepression[24].

The recent discovery of myokines — that is, cytokines that are produced and secreted by skeletal muscle cells — sheds light on the association between exercise and inflammation[20]. The first myokine to be described was IL-6; other similar factors, including IL-8 (also known as CXCL8) and IL-15, are produced and secreted when muscle fibres

[1]Institute of Physiology and Zurich Center for Integrative Human Physiology (ZIHP), University of Zurich, CH-8057 Zurich, Switzerland. [2]Dana-Farber Cancer Institute and Department of Cell Biology, Harvard Medical School, Boston, Massachusetts 02115, USA.

Box 1 | Clinical consequences of a sedentary lifestyle

Inactivity is a risk factor for many chronic disorders, regardless of age, gender, race and health status. It also affects quality of life and life expectancy. Possible clinical consequences of inactivity are listed.

Metabolic conditions
Obesity, type 2 diabetes, dyslipidaemia (dysregulated lipid metabolism) and hypercholesterolaemia, metabolic syndrome and gallstone formation

Cardiovascular diseases
Hypertension, intermittent claudication (leg pain when walking but not when resting, angina, platelet adhesion and aggregation, atherosclerosis, thrombosis, coronary artery disease, myocardial infarction (heart attack), heart failure and stroke

Pulmonary diseases
Asthma and chronic obstructive pulmonary disease

Cancers
Breast cancer, colon cancer, endometrial cancer, prostate cancer, pancreatic cancer and melanoma

Neurological disorders
Learning and memory impairment, cognitive dysfunction, dementia, depression, mood and anxiety disorders and neurodegeneration (such as occurs in Alzheimer's disease, Huntington's disease and Parkinson's disease)

Musculoskeletal disorders
Lower-back pain, osteoporosis and related fractures, osteoarthritis and rheumatoid arthritis

Gastrointestinal conditions
Reduced intestinal motility, and constipation

Immune-system alterations
Immune dysfunction and chronic inflammation

Sarcopenia
Age-related muscle wasting

Reduced quality of life
Physical frailty, decreased psychological well-being, decreased ability to carry out daily tasks and social interactions, decreased functional independence, decreased mobility, increased susceptibility to psychological stress, impaired reaction skills, and impaired sense of balance, flexibility and agility

Shorter life expectancy

contract[25]. In addition to these myokines, more IL-1 receptor antagonist, IL-10 and TNF-α are found in the blood after exercise[25]. This change in TNF-α concentration is restricted to physical activity of extremely high intensity, so TNF-α might be the factor responsible for the heightened inflammatory state that occurs after prolonged, intense exercise.

After myokines have been released transiently into the circulation, they mediate some of the beneficial systemic effects of exercise on non-muscle tissues. IL-6, for example, modulates glucose production by the liver. Some of these myokines are clearly pro-inflammatory (for example, IL-1β and TNF-α) or anti-inflammatory (for example, IL-10 and IL-1 receptor antagonist). Paradoxically, other myokines have been reported to have both pro-inflammatory effects and anti-inflammatory effects[26]. For example, having a serum IL-6 concentration that is chronically higher than a standard basal concentration in healthy individuals is a predictor of developing obesity and type 2 diabetes. In addition, chronically increased concentrations of systemic IL-1β, IL-6, IL-8, IL-10 and TNF-α have been linked to the development of many diseases associated with inflammation, including cancer and other ageing-associated disorders, such as sarcopenia, neurodegeneration and depression[10,11,13,27–29]. Finally, a chronic increase in IL-6 and TNF-α concentrations results in atrophy of skeletal muscle and inhibition of muscle regeneration, respectively[30,31]. Thus, the transient fluctuations in local and systemic myokine concentrations after physical activity might contribute to the beneficial effects of exercise on non-muscle tissues in a hormone-like manner, whereas a chronic increase in the concentration of many of these molecules is likely to be pro-inflammatory and detrimental. It is clear therefore that the increase in IL-6 and other cytokines secreted by muscle during exercise, and the subsequent return of these cytokine concentrations to basal levels, is tightly regulated.

Endurance and strength training

Distinct exercise regimens are useful for preventing and treating different pathologies. Endurance training improves cardiovascular function[32], whereas strength training reduces sarcopenia[33]. A combination of both training regimens has recently been reported to be the most beneficial for individuals with type 2 diabetes[34]. For preventing or treating other diseases, the optimal type of training remains to be defined. Interestingly, moderate exercise (such as walking) is sufficient to reduce the risk of developing dementia, as shown by a prospective study involving individuals of 65 years of age or older[35].

Strength training and endurance training activate distinct signal-transduction pathways and result in specific adaptations of skeletal muscle. The capacity of a muscle for adaptation depends mainly on the relative number and cross-sectional area of the different muscle-fibre types in a particular muscle[36]. Endurance improves with an increase in the number of type I and type IIA muscle fibres (as occurs during endurance training). Type I and type IIA muscle fibres are red in appearance and are characterized by a large number of mitochondria, more myoglobin and vascularization than type IIB and type IIX muscle fibres (which are the other main fibre types), and the presence of a specific set of myofibrillar proteins. These characteristics render the muscle fibres resistant to fatigue and enable them to contract slowly with a low peak force[37]. The main source of ATP for these fibres is oxidative phosphorylation of glucose and non-esterified fatty acids. By contrast, type IIB and type IIX muscle fibres appear white, have relatively few mitochondria and mainly metabolize phosphocreatine and glucose anaerobically to generate ATP. Both type IIB and type IIX muscle fibres are present in rodents, whereas humans have only type IIX muscle fibres. These fibres fatigue rapidly but can generate fast contractions with a high peak force[37]. When stimulated by strength training, type IIB and type IIX muscle fibres can undergo substantial hypertrophy[38].

PGC1α and skeletal muscle physiology

The adaptation of muscle fibres to training is mediated by the firing patterns of their motor neurons[39]. These neurons initiate skeletal muscle contractions, through calcium signalling pathways. The gene-expression patterns that are specific to type I and type IIA muscle fibres are achieved by the frequent release of small amounts of calcium from the sarcoplasmic reticulum into the cytosol, as is observed during endurance training. Strength training, by contrast, results in the intermittent release of large amounts of calcium from the sarcoplasmic reticulum; this promotes the transcription of genes that mediate a type-IIB-specific and type-IIX-specific muscle-fibre response (that is, a contraction after a single burst of calcium and a change in fibre type after repeated bursts of calcium), as well as fibre hypertrophy[39]. With either type of training, the release of calcium from the sarcoplasmic reticulum activates the catalytic subunit of the protein phosphatase calcineurin and members of the calcium/calmodulin-dependent protein-kinase family, which together alter the phosphorylation state of multiple transcription factors and their coactivators[40].

This heightened calcium signalling activates several important transcription factors: cyclic-AMP-responsive-element-binding protein (CREB), myocyte-enhancer factor 2C (MEF2C) and MEF2D, and

members of the nuclear factor of activated T cells (NFAT) family. These factors then alter the expression of exercise-regulated muscle genes, particularly the gene encoding the powerful transcriptional coactivator PGC1α (peroxisome-proliferator-activated receptor-γ (PPAR-γ) coactivator 1α)[41]. Accordingly, the expression of this gene, *PGC1A* (also known as *PPARGC1A*), is rapidly induced after a single bout of endurance exercise[42]. When physical activity stops, the amounts of *PGC1A* messenger RNA and PGC1α protein quickly revert to the quantities present before exercise[42]. During acute bouts of exercise, the increase in *PGC1A* expression is probably mainly a mechanism for modulating metabolic fluxes in skeletal muscle, in response to a decrease in ATP and altered fuel demands[43]. The multifaceted interaction of PGC1α with AMP-activated protein kinase is likely to be important for modulating these metabolic fluxes[44]: on activation, this protein kinase both phosphorylates PGC1α and induces the *de novo* synthesis of more PGC1α. In contrast to the transient increase in PGC1α that occurs during acute exercise, more PGC1α is present in chronically exercised skeletal muscle, even between individual bouts of exercise, than in untrained muscle[45]. This reflects the difference between short-term adaptation and long-term adaptation of skeletal muscle to endurance training.

Changes in muscle plasticity that are induced by chronic endurance exercise — such as fibre-type switching towards type I and type IIA muscle fibres, which are more oxidative and have greater endurance capacity than type IIB and type IIX muscle fibres — therefore correlate with an increase in the basal expression of *PGC1A*[45]. Furthermore, even when muscle fibres are in a rested state, more PGC1α is present in oxidative (type I and type IIA) fibres than in glycolytic (type IIB and type IIX) fibres irrespective of whether the muscles are trained[46]. The role of PGC1α has been further confirmed by studies in transgenic mice. When a transgene encoding PGC1α is expressed in skeletal muscle at levels up to those observed in type I muscle fibres, there is a stable and robust fibre-type switch towards both type I and type IIA muscle fibres[46]. Individual muscle fibres from these mice are more resistant to fatigue than are fibres from wild-type mice, and these mice perform better at endurance exercise, indicating that the chronic increase in the amount of PGC1α mediates many (if not all) of the phenotypic changes observed in muscle that has undergone endurance training[46,47]. The fibre switch promoted by PGC1α in these mice is characterized by an increase in mitochondrial density and function, an increase in oxidative metabolism, an increase in expression of myofibrillar proteins that are characteristic of type I and type IIA muscle fibres, and a switch in substrate fuel usage[46,48]. Conversely, mice in which the gene encoding PGC1α has been selectively ablated in skeletal muscle (that is, muscle-specific $Pgc1a^{-/-}$ mice) have more type IIB and type IIX (glycolytic) muscle fibres and a lower capacity for endurance exercise than wild-type mice[49]. From these studies, it is clear that PGC1α is a key mediator of many of the known beneficial effects of physical activity on skeletal muscle physiology[50,51].

PGC1α and chronic inflammation

One of the most important effects of exercise on human health is preventing protein catabolism and muscle wasting. These processes occur during limb immobilization, prolonged hospitalization, and various muscular dystrophies and other diseases. Therefore, changing the phenotypes of muscles in these patients to those of exercised muscles would improve the patients' health and overall quality of life, but these patients often cannot exercise effectively. Several studies indicate, however, that PGC1α prevents protein catabolism and muscle wasting in a variety of contexts. For example, denervation-induced muscle atrophy, muscle damage caused by treatment with statins, and the effects of Duchenne's muscular dystrophy are greatly ameliorated when the amount of PGC1α is maintained at normal levels or increased[52–54].

The precise mechanisms by which PGC1α mediates these beneficial effects are unclear, but there are several possibilities (Fig. 2). Increasing the expression of mitochondrial genes and other metabolic genes, and the resultant correction of the 'energy crisis' that is associated with muscular dystrophies[44,55], is a plausible mechanism. Other factors that are likely to contribute to the protection against muscle wasting provided by PGC1α are a decrease in the transcription of atrophy-specific genes (by inhibiting the activity of the transcription factor forkhead box O3 (FOXO3))[54], an increase in the transcriptional program for protein synthesis[54,56], and stabilization of the postsynaptic side of the junction between the motor neuron and the muscle fibre (the neuromuscular junction)[53]. In particular, regulating the genes that encode the postsynaptic components of the neuromuscular junction could ameliorate the pathology of neuromuscular disorders in which this junction has a reduced function, even those in which aberrant motor neuron function, and not skeletal muscle dysfunction, is the primary cause.

A key observation that might be relevant to a much broader set of chronic disorders arose from detailed studies of the previously mentioned muscle-specific $Pgc1a^{-/-}$ mice[49,57]. Many genes that are involved in local inflammation or systemic inflammation were transcribed in the muscles of these mice, particularly those encoding the inflammatory cytokines IL-6 and TNF-α, and suppressor of cytokine signalling 1 (SOCS1), SOCS3 and CD68 (refs 49, 57). In addition, mice with skeletal muscle cells that were heterozygous for *Pgc1a* ($Pgc1a^{+/-}$) showed a smaller, but significant, increase in the expression of many of these genes[57]. In both cases, a chronic increase in the amount of circulating

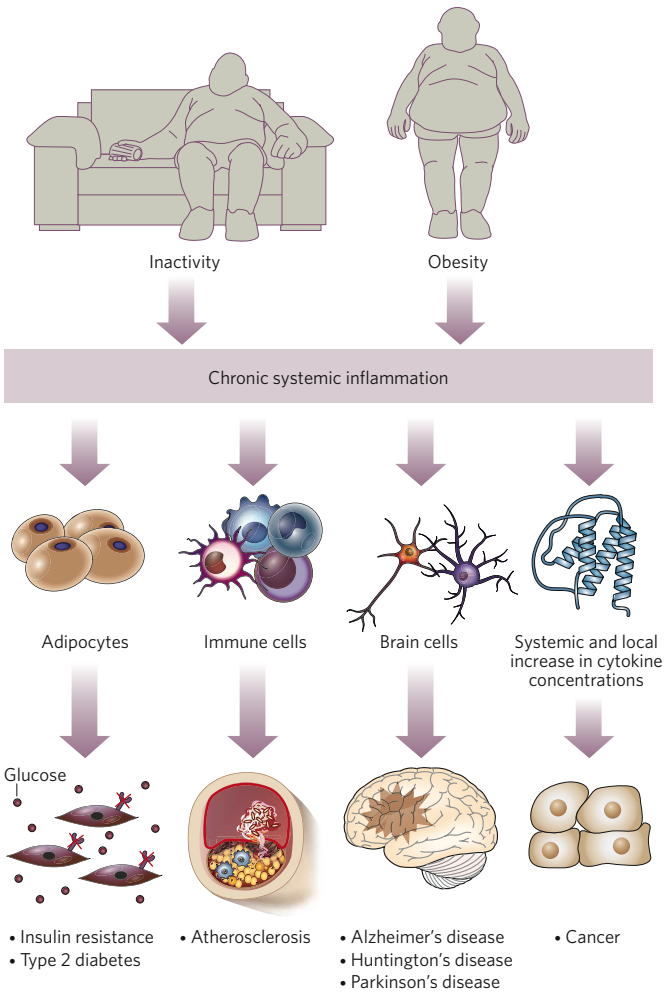


Figure 1 | Inflammation and chronic diseases. Inactivity and obesity trigger persistent, low-grade systemic inflammation. Moreover, inflammation in certain tissues is linked to the development of many chronic diseases. Examples of such tissues and the consequences of inflammation are shown. Inflammatory cytokines released from adipose tissue are linked to the development of insulin resistance and type 2 diabetes. Inflammatory responses by immune cells and glial cells are associated with atherosclerosis and neurodegenerative diseases, respectively. The systemic and local production of cytokines contributes to the aetiology of certain cancers.

IL-6 was observed[57]. Similarly, *Pgc1a*$^{-/-}$ primary muscle cells that had been differentiated into myotubes *in vitro* contained more *Tnfa* and *Il6* mRNA and secreted more IL-6 into the culture medium than did wild-type myotubes[57]. Conversely, expression of *Pgc1a* from an adenoviral vector in myotubes derived from C2C12 cells (an immortalized mouse muscle cell line) reduced the amount of *Tnfa* and *Il6* mRNA *in vitro*[57]. Taken together, these data strongly suggest that, after ablation of *Pgc1a*, at least some of the inflammatory cytokines that circulate *in vivo* originate from the skeletal muscle cells themselves, suggesting that PGC1α usually functions to suppress the production of such inflammatory mediators. Clearly, the transcriptional program by which muscle cells produce these cytokines might be amplified by the recruitment of particular immune cells, which can amplify an inflammatory response, to the muscle.

Importantly, mice with *Pgc1a*$^{+/-}$ skeletal muscle cells have a reduction in *Pgc1a* mRNA[57] that is quantitatively comparable to the transcriptional dysregulation (a 36% reduction) observed in the muscle of people with type 2 diabetes compared with healthy volunteers[58,59]. Furthermore, this reduction in expression corresponds quantitatively to the reduction in *Pgc1a* mRNA observed in the muscles of sedentary mice compared with exercising mice[54]. Although it is not possible in humans to establish whether the link between reduced expression of PGC1α and the expression of inflammatory genes is causal, people with type 2 diabetes also have increased transcription of the genes encoding IL-6 and TNF-α in skeletal muscle, as well as an increased concentration of IL-6 in serum[57]. Thus, the reduction of *PGC1A* mRNA in the skeletal muscle of people with type 2 diabetes is likely to be closely linked to the chronic, low-grade inflammation that is present in these individuals. Finally, individuals with early-stage (preclinical) type 2 diabetes also have less skeletal muscle PGC1α than healthy volunteers, probably contributing to an increase in the systemic concentration of IL-6, a strong predictor that type 2 diabetes will develop[60]. Indeed, skeletal muscle PGC1α levels correlate inversely with expression of *IL-6* and *TNFA* in individuals with normal glucose tolerance or type 2 diabetes[57]. Taken together, these findings strongly suggest that there is a causal relationship between the increases in PGC1α expression observed in human skeletal muscle after physical activity and the reduction in cytokine release from skeletal muscle that is known to occur with moderate exercise. Conversely, the effect on mice of losing even one *Pgc1a* allele, stimulating a broad program of cytokine expression and release, strongly suggests that a similar process occurs in humans who engage in chronic sedentary behaviour. For example, both muscle-specific *Pgc1a*$^{-/-}$ mice and individuals not engaging in adequate physical activity have a decreased capacity to exercise and a muscle-fibre-type switch towards glycolytic muscle fibres[49].

The molecular mechanisms that link PGC1α and inflammatory gene expression in muscle are unknown, but they might reflect the role of PGC1α in the control of reactive oxygen species (ROS). PGC1α is a powerful suppressor of ROS production, which it does in parallel with increasing mitochondrial respiration. This process occurs through the PGC1α-mediated expression of enzymes involved in ROS detoxification, as well as uncoupling proteins that can attenuate ROS production[61,62]. In fact, increased oxidative stress and inflammation are known to go hand in hand in many skeletal-muscle-associated diseases[63]. Specifically, ROS have been shown to induce inflammatory cytokine production in skeletal muscle[64]. Thus, the decreased expression of the anti-ROS genes (that is, those encoding the ROS-detoxification enzymes and the uncoupling proteins) in muscle-specific *Pgc1a*$^{-/-}$ mice[57] is likely to make a substantial contribution to the observed increases in cytokine expression. Clearly, PGC1α might have other, more-direct, effects on the expression of genes whose products have pro-inflammatory or anti-inflammatory activity.

Analysis of muscle-specific *Pgc1a*$^{-/-}$ mice revealed that dysregulation of PGC1α in skeletal muscle does not cause insulin resistance in this tissue but causes abnormal systemic glucose and insulin homeostasis as a result of reduced insulin levels and abnormal pancreatic islet morphology. This unexpected effect at a distal site seems to result from a harmful cross-talk between skeletal muscle and pancreatic β-cells in these animals[57]. An increase in systemic inflammation is one likely mechanism by which skeletal muscle with dysregulated expression of the gene encoding PGC1α could modulate β-cell function: the concentration of IL-6 in the blood of muscle-specific *Pgc1a*$^{-/-}$ mice increases concomitant with a reduction in insulin secretion by β-cells[57]. Similarly, in isolated islets *in vitro*, glucose-stimulated insulin secretion by β-cells is reduced in response to IL-6 in both wild-type mice and muscle-specific *Pgc1a*$^{-/-}$ mice[57]. These data indicate unambiguously that the amount of skeletal muscle PGC1α can markedly affect the function of pancreatic islets. Given this effect, skeletal muscle PGC1α

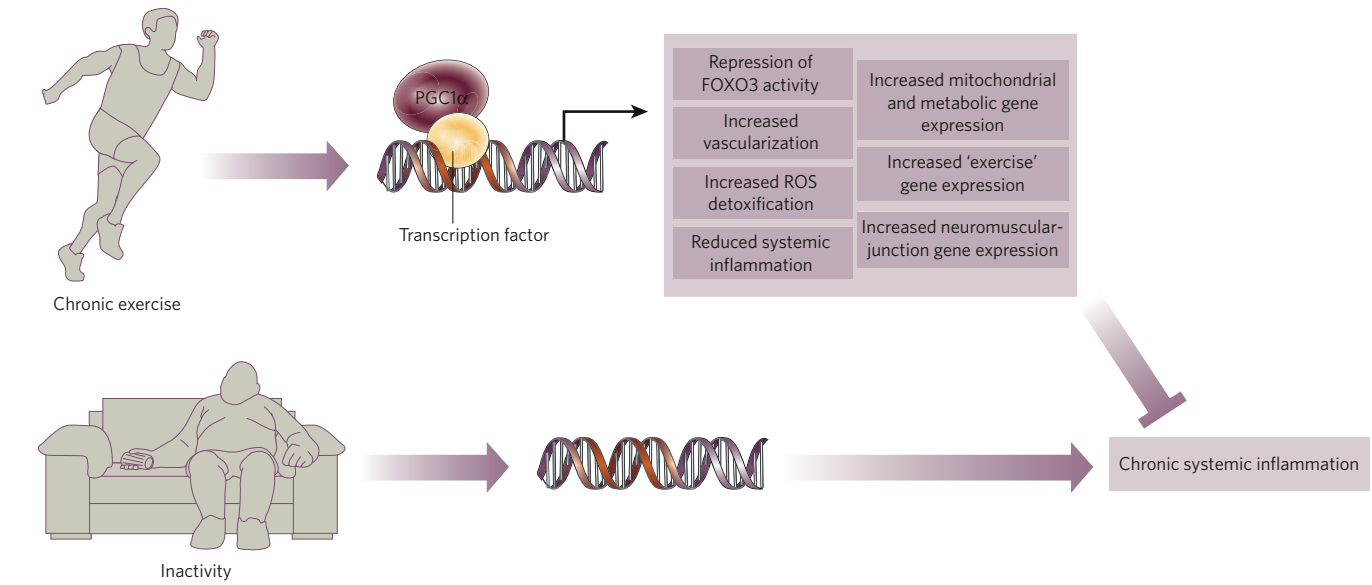


Figure 2 | Effect of PGC1α on chronic systemic inflammation. Physical activity determines the amount of PGC1α in skeletal muscle: the more activity, the more PGC1α. PGC1α, in turn, controls the adaptation of muscle fibres to exercise and confers several benefits. Consequently, a reduction in systemic inflammation is observed in individuals who exercise, particularly in those who engage in chronic exercise. By contrast, inactivity, and thus small amounts of PGC1α in skeletal muscle, results in a chronic systemic inflammatory state, which has serious pathological consequences. This inactivity-driven systemic inflammation is further exacerbated by obesity (not shown). FOXO3, forkhead box O3; ROS, reactive oxygen species.

levels probably influence the structure and functions of other tissues and organs as well.

Systemic effects of exercise and PGC1α

We suggest here that the decrease in *PGC1A* expression in skeletal muscle that results from sedentary behaviour can elicit a low-level chronic inflammatory response, which has a negative impact on many other tissues. As noted earlier, many (if not all) chronic diseases of ageing — including cardiovascular disease, cancer and neurodegeneration — are associated with chronic inflammation. In many cases, this association has been shown to be causal in defined experimental systems. Suppressing chronic inflammation in muscle through the exercise-mediated induction of *PGC1A* expression would be expected to lower the frequency and/or severity of these diseases. In terms of clinical data, exercise has many neurological benefits, most notably improved learning and memory, protection against neurodegeneration, and amelioration of depression and other mood disorders[18]. Cancer of the colon, breast, prostate, endometrium, pancreas and skin all have a greater incidence in inactive individuals than in those who exercise[1,65]. Thus, multi-organ health and plasticity as a result of exercise might be strongly affected by altering the level of systemic inflammation through controlling the amount and activity of PGC1α in skeletal muscle.

It is important to note that it is unlikely that any of the chronic diseases discussed here are caused by a reduction in PGC1α alone. Instead, these diseases are multifactorial: that is, several 'insults' are required for a chronic disease to fully develop. The multifactorial nature of human cancer is a useful conceptual model for most chronic diseases. The multiple insults required might originate from the genetic heritage of a patient, be acquired as a result of spontaneous somatic mutations, and/or result from environmental or lifestyle factors. Thus, variables such as sedentary behaviour and reduced amounts of skeletal muscle PGC1α could be viewed, on a population basis, as shifting the likelihood of developing disease compared with individuals who engage in the recommended amount of exercise or physical activity (http://www.cdc.gov/nccdphp/dnpa/physical/everyone/recommendations/index.htm) (Fig. 3). For any given individual, it is difficult to determine the impact of sedentary behaviour on the risk of developing a particular disease. But it is clear that widespread sedentary behaviour has an impact on the population overall. Conversely, populations with above-average physical activity and thus more skeletal muscle PGC1α have a lower incidence of disease than those with average activity and therefore average amounts of skeletal muscle PGC1α.

Obesity and inactivity

Sedentary behaviour often contributes to the development of obesity and is often concurrent with obesity[1,66,67]. Conversely, individuals who are obese are less likely to exercise[66,67]. Inactivity and obesity are also independent risk factors for many of the same chronic diseases. In fact, inactivity increases the risk of developing chronic diseases irrespective of body mass index (a measure of obesity). Thus, as an independent risk factor, inadequate physical activity exacerbates the detrimental effects of obesity[1]. We predict that lack of exercise and obesity interact to have a harmful effect on the transcriptional programs discussed here.

Obesity has been strongly associated with an inflammatory transcriptional program, which includes the expression of genes encoding TNF-α, IL-1β and IL-6 (ref. 10). More precisely, it has been known for more than a decade that adipose tissue, in the context of obesity, secretes increased amounts of these cytokines in both rodents and humans[68,69]. Moreover, TNF-α and other adipokines (cytokines produced by adipocytes) have been shown to have a functional role in the insulin resistance that is observed in obese mice[10]. If it is assumed that there is a quantitative threshold of systemic cytokines that needs to be present chronically for pathology to occur in other tissues (that is, non-adipose and non-muscle tissues), then this threshold is much more likely to be reached in individuals who are both inactive and obese than in those who are just inactive (Fig. 3). Furthermore, if the age of onset of a particular disease or the extent of disease is dose responsive with respect to the concentrations

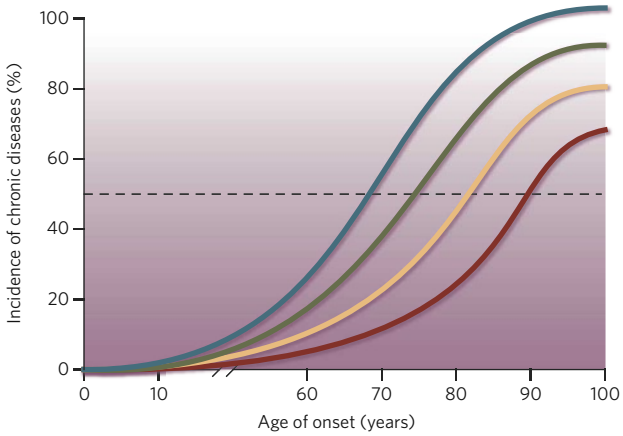


Figure 3 | Inactivity and obesity as risk factors for developing chronic diseases. How lifestyle affects the age of onset of chronic diseases and the incidence of these diseases is depicted, using theoretical estimates. The baseline is individuals who engage in the recommended amount of physical activity (yellow). A sedentary lifestyle (green) lowers the threshold for the age of onset and the incidence of chronic diseases (see dashed line). Inactivity and obesity together (blue) further increase the risk for developing chronic diseases. Those who exercise chronically (red) have a reduced risk.

of systemic inflammatory molecules, then sedentary behaviour coupled with obesity would be expected to bring about disease earlier and/or in a more severe form[70]. Obesity and sedentary behaviour might also interact in ways that are not just quantitative. Complete lists of the cytokines that contribute to the onset of particular diseases are not yet known, but obesity and sedentary behaviour might bring together a particular combination of adipokines and myokines that function additively or even synergistically to cause disease. But how large are the possible synergistic effects of obesity and sedentary behaviour in humans? Similar to other modifiable factors — such as smoking, type 2 diabetes and hypertension — obesity is predicted to reduce an individual's life expectancy by 1–5 years. By contrast, physical activity is estimated to increase it by up to 5 years. Importantly, a composite lifestyle of healthy behaviours — for example, engaging in sufficient exercise while having a healthy diet — has been proposed to add 10 years to the average life expectancy[8,71].

Testing the hypothesis

It is clear from the findings presented here that many aspects of the physiological and pathophysiological effects of modulating PGC1α in skeletal muscle and other organs remain enigmatic. Even less is known about the functions and therapeutic potential of the other two members of this gene family: the genes encoding PGC1β and PGC1-related coactivator (PRC). Similar to PGC1α, transgenic expression of PGC1β in skeletal muscle increases the capacity of mice for endurance exercise in comparison to wild-type mice; however, different mechanisms seem to control the PGC1α- and PGC1β-mediated increases in endurance[72].

In the future, altering the amount and/or activity of PGC1α in skeletal muscle might be a useful way to prevent and treat certain diseases. The effects of musculoskeletal disorders — disuse-induced muscle atrophy[54], Duchenne's muscular dystrophy[53] and statin-mediated muscle wasting[52] — have already been ameliorated by altering PGC1α in this manner in experimental animal models. The potential of skeletal muscle PGC1α to modulate non-muscular diseases has not yet been studied, but the ideas that we have presented are readily testable in experimental animal models. For example, we propose that mice lacking one or both copies of *Pgc1a* in skeletal muscle are more susceptible to cancers, cardiovascular disease and neurological disorders. To test this idea, mice lacking *Pgc1a* could be treated with chemical carcinogens that cause cancer of the breast, colon or other tissues, and the rate of tumour formation and progression could be determined quantitatively. Similarly, these mice could be given challenges that induce heart failure or

certain types of neurodegeneration that model Parkinson's disease or Alzheimer's disease, and the rates of incidence and progression could be monitored.

To improve human health, it will be important to determine appropriate exercise regimens that will protect against diseases associated with muscle-based inflammation. But chemical modulation of the PGC1α pathway in skeletal muscle could also have a huge impact in the clinic. Drugs and drug-like compounds have been shown to modulate the expression of *Pgc1a* in mouse muscle cells *in vitro*[56,73,74]. In addition, several of the transcription factors that bind to PGC1α and thus regulate transcription in skeletal muscle have been identified[50,51] and could be therapeutic targets. For example, when PGC1α is recruited by the transcription factor oestrogen-related receptor-α, the mitochondrial genes involved in oxidative phosphorylation are transcribed, and these genes are dysregulated in the muscle of people with type 2 diabetes[56]. Pharmacological disruption of the interaction between these two proteins in cultured muscle cells induces a metabolic phenotype resembling that of *in vivo* muscle in an individual with type 2 diabetes[56]. These findings point to how PGC1α could be targeted selectively[75], an important consideration if PGC1α is to be modulated in the clinic. If successful, therapeutic modulation of PGC1α has a huge potential for the treatment of patients with muscle wasting, sarcopenia, type 2 diabetes and muscular dystrophies, as well as other serious non-muscular chronic diseases. ■

1. Booth, F. W., Chakravarthy, M. V., Gordon, S. E. & Spangenburg, E. E. Waging war on physical inactivity: using modern molecular ammunition against an ancient enemy. *J. Appl. Physiol.* **93,** 3–30 (2002).
2. Eriksson, G. *et al.* Changes in physical fitness and changes in mortality. *Lancet* **352,** 759–762 (1998).
3. Hu, F. B. *et al.* Adiposity as compared with physical activity in predicting mortality among women. *N. Engl. J. Med.* **351,** 2694–2703 (2004).
4. Kokkinos, P. *et al.* Exercise capacity and mortality in black and white men. *Circulation* **117,** 614–622 (2008).
5. Booth, F. W. & Lees, S. J. Fundamental questions about genes, inactivity, and chronic diseases. *Physiol. Genom.* **28,** 146–157 (2007).
6. McCracken, M., Jiles, R. & Blanck, H. M. Health behaviors of the young adult U.S. population: behavioral risk factor surveillance system, 2003. *Prevent. Chron. Dis.* **4,** A25 (2007).
7. Hollmann, W., Struder, H. K., Tagarakis, C. V. & King, G. Physical activity and the elderly. *Eur. J. Cardiovasc. Prev. Rehabil.* **14,** 730–739 (2007).
8. Yates, L. B. *et al.* Exceptional longevity in men: modifiable factors associated with survival and function to age 90 years. *Arch. Intern. Med.* **168,** 284–290 (2008).
9. Knowler, W. C. *et al.* Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *N. Engl. J. Med.* **346,** 393–403 (2002).
10. Hotamisligil, G. S. Inflammation and metabolic disorders. *Nature* **444,** 860–867 (2006).
11. Haffner, S. M. The metabolic syndrome: inflammation, diabetes mellitus, and cardiovascular disease. *Am. J. Cardiol.* **97,** 3A–11A (2006).
12. Matter, C. M. & Handschin, C. RANTES (regulated on activation, normal T cell expressed and secreted), inflammation, obesity, and the metabolic syndrome. *Circulation* **115,** 946–948 (2007).
13. Lin, W. W. & Karin, M. A cytokine-mediated link between innate immunity, inflammation, and cancer. *J. Clin. Invest.* **117,** 1175–1183 (2007).
14. Zhou, J. R., Blackburn, G. L. & Walker, W. A. Symposium introduction: metabolic syndrome and the onset of cancer. *Am. J. Clin. Nutr.* **86,** S817–S819 (2007).
15. Tansey, M. G. *et al.* Neuroinflammation in Parkinson's disease: is there sufficient evidence for mechanism-based interventional therapy? *Front. Biosci.* **13,** 709–717 (2008).
16. Whitton, P. S. Inflammation as a causative factor in the aetiology of Parkinson's disease. *Br. J. Pharmacol.* **150,** 963–976 (2007).
17. Zipp, F. & Aktas, O. The brain as a target of inflammation: common pathways link inflammatory and neurodegenerative diseases. *Trends Neurosci.* **29,** 518–527 (2006).
18. Cotman, C. W., Berchtold, N. C. & Christie, L. A. Exercise builds brain health: key roles of growth factor cascades and inflammation. *Trends Neurosci.* **30,** 464–472 (2007).
19. Perry, V. H., Cunningham, C. & Holmes, C. Systemic infections and inflammation affect chronic neurodegeneration. *Nature Rev. Immunol.* **7,** 161–167 (2007).
20. Febbraio, M. A. Exercise and inflammation. *J. Appl. Physiol.* **103,** 376–377 (2007).
21. Gleeson, M. Immune function in sport and exercise. *J. Appl. Physiol.* **103,** 693–699 (2007).
22. Nieman, D. C. Current perspective on exercise immunology. *Curr. Sports Med. Rep.* **2,** 239–242 (2003).
23. Gleeson, M., McFarlin, B. & Flynn, M. Exercise and Toll-like receptors. *Exerc. Immunol. Rev.* **12,** 34–53 (2006).
24. Gleeson, M., Nieman, D. C. & Pedersen, B. K. Exercise, nutrition and immune function. *J. Sports Sci.* **22,** 115–125 (2004).
25. Pedersen, B. K., Akerstrom, T. C., Nielsen, A. R. & Fischer, C. P. Role of myokines in exercise and metabolism. *J. Appl. Physiol.* **103,** 1093–1098 (2007).
26. Kristiansen, O. P. & Mandrup-Poulsen, T. Interleukin-6 and diabetes: the good, the bad, or the indifferent? *Diabetes* **54,** S114–S124 (2005).
27. Sarkar, D. & Fisher, P. B. Molecular mechanisms of aging-associated inflammation. *Cancer Lett.* **236,** 13–23 (2006).
28. Bremmer, M. A. *et al.* Inflammatory markers in late-life depression: Results from a population-based study. *J. Affect. Disord.* **106,** 249–255 (2008).
29. Roubenoff, R. Physical activity, inflammation, and muscle loss. *Nutr. Rev.* **65,** S208–S212 (2007).
30. Haddad, F., Zaldivar, F., Cooper, D. M. & Adams, G. R. IL-6-induced skeletal muscle atrophy. *J. Appl. Physiol.* **98,** 911–917 (2005).
31. Coletti, D. *et al.* Tumor necrosis factor-α gene transfer induces cachexia and inhibits muscle regeneration. *Genesis* **43,** 120–128 (2005).
32. Manson, J. E. *et al.* A prospective study of walking as compared with vigorous exercise in the prevention of coronary heart disease in women. *N. Engl. J. Med.* **341,** 650–658 (1999).
33. Thomas, D. R. Loss of skeletal muscle mass in aging: examining the relationship of starvation, sarcopenia and cachexia. *Clin. Nutr.* **26,** 389–399 (2007).
34. Sigal, R. J. *et al.* Effects of aerobic training, resistance training, or both on glycemic control in type 2 diabetes: a randomized trial. *Ann. Int. Med.* **147,** 357–369 (2007).
35. Larson, E. B. *et al.* Exercise is associated with reduced risk for incident dementia among persons 65 years of age and older. *Ann. Int. Med.* **144,** 73–81 (2006).
36. Pette, D. Historical perspectives: plasticity of mammalian skeletal muscle. *J. Appl. Physiol.* **90,** 1119–1124 (2001).
37. Flück, M. & Hoppeler, H. Molecular basis of skeletal muscle plasticity — from gene to form and function. *Rev. Physiol. Biochem. Pharmacol.* **146,** 159–216 (2003).
38. Glass, D. J. Skeletal muscle hypertrophy and atrophy signaling pathways. *Int. J. Biochem. Cell Biol.* **37,** 1974–1984 (2005).
39. Chin, E. R. *et al.* A calcineurin-dependent transcriptional pathway controls skeletal muscle fiber type. *Genes Dev.* **12,** 2499–2509 (1998).
40. Berchtold, M. W., Brinkmeier, H. & Muntener, M. Calcium ion in skeletal muscle: its crucial role for muscle function, plasticity, and disease. *Physiol. Rev.* **80,** 1215–1265 (2000).
41. Puigserver, P. *et al.* A cold-inducible coactivator of nuclear receptors linked to adaptive thermogenesis. *Cell* **92,** 829–839 (1998).
42. Pilegaard, H., Saltin, B. & Neufer, P. D. Exercise induces transient transcriptional activation of the PGC-1α gene in human skeletal muscle. *J. Physiol.* **546,** 851–858 (2003).
43. Hood, D. A., Irrcher, I., Ljubicic, V. & Joseph, A. M. Coordination of metabolic plasticity in skeletal muscle. *J. Exp. Biol.* **209,** 2265–2275 (2006).
44. Jager, S., Handschin, C., St-Pierre, J. & Spiegelman, B. M. AMP-activated protein kinase (AMPK) action in skeletal muscle via direct phosphorylation of PGC-1α. *Proc. Natl Acad. Sci. USA* **104,** 12017–12022 (2007).
45. Russell, A. P. *et al.* Endurance training in humans leads to fiber type-specific increases in levels of peroxisome proliferator-activated receptor-γ coactivator-1 and peroxisome proliferator-activated receptor-α in skeletal muscle. *Diabetes* **52,** 2874–2881 (2003).
46. Lin, J. *et al.* Transcriptional co-activator PGC-1α drives the formation of slow-twitch muscle fibres. *Nature* **418,** 797–801 (2002).
47. Calvo, J. A. *et al.* Muscle-specific expression of PPARγ coactivator-1α improves exercise performance and increases peak oxygen uptake. *J. Appl. Physiol.* **104,** 1304–1312 (2008).
48. Wende, A. R. *et al.* A role for the transcriptional coactivator PGC-1α in muscle refueling. *J. Biol. Chem.* **282,** 36642–36651 (2007).
49. Handschin, C. *et al.* Skeletal muscle fiber-type switching, exercise intolerance, and myopathy in PGC-1α muscle-specific knock-out animals. *J. Biol. Chem.* **282,** 30014–30021 (2007).
50. Handschin, C. & Spiegelman, B. M. PGC-1 coactivators, energy homeostasis, and metabolism. *Endocr. Rev.* **27,** 728–735 (2006).
51. Lin, J., Handschin, C. & Spiegelman, B. M. Metabolic control through the PGC-1 family of transcription coactivators. *Cell Metab.* **1,** 361–370 (2005).
52. Hanai, J. I. *et al.* The muscle-specific ubiquitin ligase atrogin-1/MAFbx mediates statin-induced muscle toxicity. *J. Clin. Invest.* **117,** 3940–3951 (2007).
53. Handschin, C. *et al.* PGC-1α regulates the neuromuscular junction program and ameliorates Duchenne muscular dystrophy. *Genes Dev.* **21,** 770–783 (2007).
54. Sandri, M. *et al.* PGC-1α protects skeletal muscle from atrophy by suppressing FoxO3 action and atrophy-specific gene transcription. *Proc. Natl Acad. Sci. USA* **103,** 16260–16265 (2006).
55. Wu, Z. *et al.* Mechanisms controlling mitochondrial biogenesis and respiration through the thermogenic coactivator PGC-1. *Cell* **98,** 115–124 (1999).
56. Mootha, V. K. *et al.* Errα and Gabpa/b specify PGC-1α-dependent oxidative phosphorylation gene expression that is altered in diabetic muscle. *Proc. Natl Acad. Sci. USA* **101,** 6570–6575 (2004).
57. Handschin, C. *et al.* Abnormal glucose homeostasis in skeletal muscle-specific PGC-1α knockout mice reveals skeletal muscle-pancreatic β cell crosstalk. *J. Clin. Invest.* **117,** 3463–3474 (2007).
58. Mootha, V. K. *et al.* PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genet.* **34,** 267–273 (2003).
59. Patti, M. E. *et al.* Coordinated reduction of genes of oxidative metabolism in humans with insulin resistance and diabetes: Potential role of PGC1 and NRF1. *Proc. Natl Acad. Sci. USA* **100,** 8466–8471 (2003).
60. Alexandraki, K. *et al.* Inflammatory process in type 2 diabetes: The role of cytokines. *Ann. NY Acad. Sci.* **1084,** 89–117 (2006).
61. St-Pierre, J. *et al.* Suppression of reactive oxygen species and neurodegeneration by the PGC-1 transcriptional coactivators. *Cell* **127,** 397–408 (2006).
62. Valle, I. *et al.* PGC-1α regulates the mitochondrial antioxidant defense system in vascular endothelial cells. *Cardiovasc. Res.* **66,** 562–573 (2005).
63. Moylan, J. S. & Reid, M. B. Oxidative stress, chronic disease, and muscle wasting. *Muscle Nerve* **35,** 411–429 (2007).
64. Ji, L. L. Modulation of skeletal muscle antioxidant defense by exercise: Role of redox signaling. *Free Radic. Biol. Med.* **44,** 142–152 (2008).

65. Brown, W. J., Burton, N. W. & Rowan, P. J. Updating the evidence on physical activity and health in women. *Am. J. Prev. Med.* **33,** 404–411 (2007).
66. Perusse, L. & Bouchard, C. Genotype-environment interaction in human obesity. *Nutr. Rev.* **57,** S31–38 (1999).
67. Rippe, J. M. & Hess, S. The role of physical activity in the prevention and management of obesity. *J. Am. Diet. Assoc.* **98,** S31–38 (1998).
68. Hotamisligil, G. S. & Spiegelman, B. M. Tumor necrosis factor α: a key component of the obesity-diabetes link. *Diabetes* **43,** 1271–1278 (1994).
69. Hotamisligil, G. S., Shargill, N. S. & Spiegelman, B. M. Adipose expression of tumor necrosis factor-α: direct role in obesity-linked insulin resistance. *Science* **259,** 87–91 (1993).
70. Hamilton, M. T., Hamilton, D. G. & Zderic, T. W. Role of low energy expenditure and sitting in obesity, metabolic syndrome, type 2 diabetes, and cardiovascular disease. *Diabetes* **56,** 2655–2667 (2007).
71. Fraser, G. E. & Shavlik, D. J. Ten years of life: Is it a matter of choice? *Arch. Intern. Med.* **161,** 1645–1652 (2001).
72. Arany, Z. *et al.* The transcriptional coactivator PGC-1β drives the formation of oxidative type IIX fibers in skeletal muscle. *Cell Metab.* **5,** 35–46 (2007).
73. Wagner, B. K. *et al.* Large-scale chemical dissection of mitochondrial function. *Nature Biotechnol.* **26,** 343–351 (2008).
74. Arany, Z. *et al.* Gene expression-based screening identifies microtubule inhibitors as inducers of PGC-1α and oxidative phosphorylation. *Proc. Natl Acad. Sci. USA* **105,** 4721–4726 (2008).
75. Handschin, C. & Mootha, V. K. Estrogen-related receptor α (ERRα): a novel target in type 2 diabetes. *Drug Discov. Today Ther. Strateg.* **2,** 151–156 (2005).

Acknowledgements We thank E. Smith for assistance with graphics. We also thank our colleagues and the members of our laboratories for comments on the manuscript, in particular S. Loffredo, J. Estall, Z. Arany, G. Hansson, S. Summermatter, M. Toigo and U. A. Meyer. C.H. is supported by the University Research Priority Program 'Integrative Human Physiology' of the University of Zurich, an SNSF-Professorship of the Swiss National Science Foundation and the Muscular Dystrophy Association. B.M.S. is supported by several grants from the National Institutes of Health.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence should be addressed to the authors (christoph.handschin@access.uzh.ch; bruce_spiegelman@dfci.harvard.edu).

NATURE|Vol 454|24 July 2008|doi:10.1038/nature07202

Integration of metabolism and inflammation by lipid-activated nuclear receptors

Steven J. Bensinger[1] & Peter Tontonoz[1]

The nuclear receptors known as PPARs and LXRs are lipid-activated transcription factors that have emerged as key regulators of lipid metabolism and inflammation. PPARs and LXRs are activated by non-esterified fatty acids and cholesterol metabolites, respectively, and both exert positive and negative control over the expression of a range of metabolic and inflammatory genes. The ability of these nuclear receptors to integrate metabolic and inflammatory signalling makes them attractive targets for intervention in human metabolic diseases, such as atherosclerosis and type 2 diabetes, as well as for the modulation of inflammation and immune responses.

The nuclear-receptor superfamily has diverse and important roles in regulating developmental, reproductive, homeostatic, inflammatory, immune and metabolic processes[1–3]. Nuclear receptors have a highly conserved structure comprising an amino-terminal domain that contains a ligand-independent activation function (AF1), a zinc-finger DNA-binding domain, a carboxy-terminal domain that binds to ligand, and a ligand-dependent transcriptional activation function (AF2)[4,5].

There are three broad classes of nuclear receptor[6]. The first class contains the prototypical (classic) ligand-driven receptors exemplified by the steroid hormone receptors, such as the oestrogen and glucocorticoid receptors. These are typically cytoplasmic, and ligand binding (hormone or steroid) induces activation, translocation to the nucleus and transcription of target genes. The second class of nuclear receptors, termed orphan receptors, encompasses a diverse set of receptors for which regulatory ligands are either undefined or do not seem to be required.

The third class of nuclear receptors comprises metabolite-activated transcription factors that form obligate heterodimers with the retinoid X receptor (RXR). In the absence of ligand, most RXR heterodimers are bound to DNA in association with co-repressors, histone deacetylases and chromatin-modifying factors to maintain active repression of target genes[7]. Ligand binding initiates a conformational change in the receptor, the exchange of co-repressors for coactivators, and the initiation of target-gene transcription. Several RXR heterodimers can both activate and repress gene expression in a signal-specific and gene-specific manner. Examples of these receptors are the liver X receptors (LXRs), the retinoic acid receptors (RARs) and the peroxisome-proliferator-activated receptors (PPARs). LXRs and PPARs have emerged as important regulators of metabolic and inflammatory signalling, particularly in metabolic disease and immunity[5].

Here we review the biology of these receptors and highlight recent work that has advanced our understanding of these receptors in physiology and pathophysiology. Current work on the dual roles of PPARs and LXRs in the control of metabolism and inflammation is likely to expand the potential therapeutic applications of their agonists.

PPARs

The PPAR family is composed of three proteins: PPAR-α, PPAR-δ (also known as PPAR-β) and PPAR-γ. (These are also known as NR1C1, NR1C2 and NR1C3, respectively.) The three PPARs have different tissue distributions and seem to have distinct but overlapping biological functions[5,8].

PPAR-α was identified in the early 1990s on the basis of it being a target of the hypolipidaemic fibrate drugs and other compounds that induce peroxisome proliferation in rodents[9]. The genes encoding PPAR-δ and PPAR-γ were cloned subsequently on the basis of both sequence homology and their function in the control of gene expression in adipose tissue[10,11]. Several studies have identified the endogenous ligands of PPARs to be unsaturated fatty acids, eicosanoids, components of oxidized low-density lipoproteins (LDLs) and very-low-density lipoproteins (VLDLs), and derivatives of linoleic acic[4,5]. In addition, a range of pharmacological ligands have been generated and are used in research and in the clinic, including the fibrates and thiazolidinediones. A more comprehensive list can be found in Table 1. Most PPAR–RXR heterodimers seem to be constitutively bound to specific response elements, known as PPAR response elements (PPREs), present in the promoters of their target genes. Most PPREs are degenerate variants of a direct repeat of the hormone response-element sequence (AGGTCA) separated by one base pair[12]. Ligand binding initiates a conformational change that results in the degradation of co-repressor complexes, the recruitment of coactivator complexes, and the subsequent induction of target-gene expression[7]. The PPAR family has been shown to modulate various cellular functions, including adipocyte differentiation, fatty-acid oxidation and glucose metabolism. Examples of target genes are discussed in this section and are reviewed in ref. 13.

Another important function of PPARs is the inhibition of inflammatory gene expression. In several model systems, PPARs repress target genes of the transcription factors nuclear factor-κB (NF-κB), nuclear factor of activated T cells (NFAT), activator protein 1 (AP1) and signal transducers and activators of transcription (STATs) in a signal-specific manner[6]. However, it has been difficult to identify a unified mechanism of repression by activated PPARs. It is possible that there are mechanisms specific to the signal, cell type, PPAR type or isoform, and even the gene promoter. Details of some of these mechanisms are discussed in this section. A further complicating issue in elucidating the mechanism of bona fide signal-specific repression is that some PPAR ligands have receptor-independent anti-inflammatory activities. Thus, it will be important to re-evaluate studies that use PPAR ligands at supraphysiological concentrations or ligands that have receptor-independent effects, such as 15-deoxy-$\Delta^{12,14}$-prostaglandin J_2.

[1]Howard Hughes Medical Institute, Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, 675 Charles E. Young Drive, Los Angeles, California 90049, USA.

Biology of PPAR-γ

Two distinct isoforms of PPAR-γ (PPAR-γ1 and PPAR-γ2) are derived from the same gene through alternative promoter usage or messenger RNA splicing. PPAR-γ has a restricted expression pattern, with prominent expression in brown and white adipose tissue, the colon, differentiated myeloid cells and the placenta[14]. Target genes of PPAR-γ include the adipocyte fatty-acid-binding protein, phosphoenolpyruvate carboxykinase, lipoprotein lipase, the uncoupling protein UCP1, the scavenger receptor CD36 and the nuclear receptor LXR-α[5]. PPAR-γ has a crucial role in the control of adipocyte differentiation that cannot be compensated for by PPAR-α or PPAR-δ, and is essential for the formation of adipose tissue *in vivo*[15,16].

A range of naturally occurring ligands can activate PPAR-γ, including unsaturated fatty acids, eicosanoids and components of oxidized LDLs (Table 1). However, the affinity of the receptor for many of these ligands is low, and in some cases the physiological relevance of the ligand has yet to be determined. PPAR-γ is the molecular target of thiazolidinediones, which sensitize cells to insulin and are in use for their antidiabetic effects in the liver, adipose tissue and skeletal muscle[17]. The mechanism of action underlying the insulin-sensitizing effects of thiazolidinediones has yet to be definitively determined and remains an area of intense research.

Table 1 | Endogenous and synthetic ligands for PPARs and LXRs

Receptor	Endogenous ligands	Synthetic ligands
PPARs		
PPAR-α	Unsaturated fatty acids, saturated fatty acids, leukotriene B_4, 8-HETE*	Clofibrate, fenofibrate, gemfibrozil, GW7647, WY14643
PPAR-δ	Unsaturated fatty acids, saturated fatty acids, carbaprostacyclin, VLDLs*	GW501516, L-165041
PPAR-γ	Unsaturated fatty acids, 15-deoxy-$\Delta^{12,14}$-prostaglandin J_2, 15-HETE, 9-HODE and 13-HODE, oxidized LDLs*	Rosiglitazone, pioglitazone, troglitazone, ciglitazone, tyrosine derivatives, farglitazar, GW7845
PPAR-α and PPAR-γ	None selective for PPAR-α and PPAR-γ only	Muraglitazar, ragaglitazar, tesaglitazar
LXRs		
LXR-α and LXR-β	22-(*R*)-hydroxycholesterol, 24-(*S*)-hydroxycholesterol, 27-hydroxycholesterol, 24-(*S*),25-epoxycholesterol	GW3965, T091317
LXR-β	None selective for LXR-β	*N*-Acylthiadiazolines

*PPAR ligands can be derived from components of oxidized LDLs and VLDLs. HETE, hydroxyeicosatetraenoic acid; HODE, hydroxyoctadecadienoic acid.

PPAR-γ and transrepression

In addition to transcriptional activation, an important function of PPAR-γ is the negative regulation of gene expression. In the unliganded state, PPAR-γ–RXR heterodimers actively repress target-gene expression. Paradoxically, deletion of the gene encoding PPAR-γ in certain contexts results in higher basal expression of PPAR-γ target genes, a phenomenon known as derepression[6,7,18]. Another repressive function of PPAR-γ is to inhibit inflammatory gene expression in a signal-specific manner. This process is termed transrepression because inhibition does not depend on the binding of PPAR–RXR heterodimers to PPREs in target-gene promoters. The mechanism of ligand-dependent transrepression remains enigmatic. Previous studies have reported that PPARs inhibit inflammatory gene expression by several mechanisms, including direct interactions with AP1 and NF-κB (thereby interfering with DNA binding by these transcription factors)[19–21], nucleocytoplasmic redistribution of the p65 subunit of NF-κB[22], modulation of p38 mitogen-activated protein-kinase activity[23], competition for limiting pools of coactivators[24], and interactions with transcriptional co-repressors[25].

A study by Christopher Glass and colleagues has made considerable advances in determining the molecular events that underlie PPAR-mediated transrepression in macrophages[26] (Fig. 1). In the basal (quiescent) state, inflammatory genes such as those encoding inducible nitric oxide synthase (iNOS; also known as NOS2) and the chemokine CC-chemokine ligand 2 (CCL2) are repressed by a protein complex containing histone deacetylases (HDACs), transducin-β-like proteins and either the nuclear-receptor co-repressor (NCOR) or the silencing mediator of retinoic acid and thyroid-hormone receptor (SMRT). These protein complexes bind to the promoters of inflammatory genes and prevent the acetylation of histones and the aggregation of coactivator complexes. The binding of lipopolysaccharide (LPS) to Toll-like receptor 4 (TLR4) at the cell surface initiates a signalling cascade that results in the ubiquitylation and subsequent degradation of the co-repressor complex by the 19S proteasome, and the translocation of NF-κB to the nucleus. The binding of NF-κB to sequence-specific elements in inflammatory gene promoters, in association with the binding of coactivator complexes to NF-κB, subsequently drives target-gene expression. In the model proposed by Glass and colleagues[26], activation of PPAR-γ inhibits inflammatory gene expression by preventing the inflammatory signal-specific removal of the co-repressor complex. Ligand activation of PPAR-γ results in a conformational change allowing the SUMOylation of PPAR-γ in the ligand-binding domain at lysine 365, mediated by the E2 ligase UBC9 and the E3 ligase PIAS1. In a manner that remains unclear, SUMOylated PPAR-γ binds to the NCOR–HDAC-containing co-repressor complex and blocks its degradation by the 19S proteasome, thereby preserving the repressed state. Future studies will address the importance of this and other proposed mechanisms of transrepression in inflammatory signalling and disease *in vivo*.

PPAR-γ in inflammation and metabolic disease

Numerous studies have shown that administration of PPAR-γ ligands can ameliorate inflammatory responses in the pancreas, lungs, joints, nervous system and gastrointestinal tract, suggesting a therapeutic potential for PPAR-γ agonists that selectively modulate inflammation (see ref. 27 for a review). For example, conditional deletion of the gene encoding PPAR-γ in macrophages and intestinal epithelial cells confirmed that PPAR-γ is important in the regulation of inflammatory bowel disease[28,29]. Ronald Evans and colleagues have also reported that endothelial-specific deletion of the gene encoding PPAR-γ showed an unexpected and important role for this protein in modulating the production of inflammatory lipids in milk in the mammary glands[30]. The ingestion of 'toxic' milk by nursing neonates resulted in growth retardation and inflammatory skin conditions. Surprisingly, macrophage-specific deletion of the gene encoding PPAR-γ was shown to regulate diet-induced obesity and insulin sensitivity, key components of type 2 diabetes and metabolic syndrome[31]. These same studies also indicated that the interleukin 4 (IL-4)–STAT6–PPAR-γ signalling axis, first described by Glass and colleagues[32], has a crucial role in macrophage differentiation and innate immunity to the protozoan parasite *Leishmania major*.

Macrophages are specialized cells that have vital roles in phagocytosis, host defence and wound healing through the regulated expression of inflammatory mediators. Recently, it has become increasingly clear that inflammation that is mediated by tissue-resident macrophages is an essential component of many metabolic, endocrine and cardiovascular disorders[33,34]. Indeed, the amelioration of inflammation leads to measurable differences in the outcome of various biological processes, including peripheral insulin resistance and atherosclerotic plaque formation. A considerable conceptual advance in our understanding of metabolic and cardiovascular disease has been the identification of two distinct functional subsets of macrophages[35]. Classically activated (M1) macrophages produce a wide variety of pro-inflammatory cytokines, have considerable antimicrobial activity and are probably involved in antigen presentation to lymphocytes. By contrast, alternatively activated (M2) macrophages generate anti-inflammatory products, mediate tissue repair and have antiparasite activity. The M1 and M2 macrophage subsets arise from the same monocyte precursors in a system that is reminiscent of the T_H1/T_H2 paradigm of T-helper-cell differentiation. The balance between these two

subsets has a crucial role in regulating inflammation in a lesion or tissue. In turn, the interplay of the M1 and M2 macrophages seems to affect the progression of various conditions and diseases, including obesity, peripheral insulin resistance, atherosclerosis and infection.

Interestingly, PPAR-γ has emerged as a crucial transcriptional regulator of monocyte phenotypic differentiation (Fig. 2). Glass and colleagues provided the initial evidence that IL-4-mediated signalling upregulates the expression of PPAR-γ and its endogenous ligands[32]. Conversely, TLR signals that promote an M1 macrophage phenotype downregulated PPAR-γ expression. More recently, Ajay Chawla and colleagues implicated PPAR-γ signalling in macrophage differentiation: they reported that the PPAR-γ coactivator 1β (PGC1β) was crucial for alternative activation[36]. Subsequent studies by this group confirmed an important role for PPAR-γ in macrophage differentiation.

Disruption of *Pparg* (the gene encoding PPAR-γ) in monocytes showed that acquisition of the alternatively activated (M2) macrophage phenotype depends on intrinsic PPAR-γ expression[31]. In the absence of PPAR-γ signalling, macrophages neither appropriately suppress inflammatory cytokine production nor acquire an oxidative metabolic program that is associated with the M2 macrophage phenotype. The functional consequences of macrophage-specific deletion of *Pparg* in BALB/c mice are increased insulin resistance in skeletal muscle and liver, and exacerbation of diet-induced obesity, two key components of metabolic syndrome[31]. Selective inactivation of PPAR-γ in monocytes also resulted in increased immunity to cutaneous leishmaniasis, a parasitic infection whose clearance depends on macrophage function. These data imply that therapeutic manipulation of PPAR-γ in monocytes might modulate immunity and infection in humans. The importance of PPAR-γ in regulating the M1/M2 macrophage switch was reconfirmed by recent work by Amine Bouhlel *et al.*[37] showing that activation of PPAR-γ potentiates polarization of circulating monocytes into macrophages of the M2 phenotype.

PPAR-γ also regulates the maturation and function of an antigen-presenting cell that is closely related to macrophages — the dendritic cell — in both mice and humans. PPAR-γ gain-of-function studies have shown that PPAR-γ signalling affects phagocytosis, cytokine production and antigen presentation[38–40]. Selective deletion of the gene encoding PPAR-γ in dendritic cells confirmed[41] the results of these studies[38–40]. Interestingly, gene-expression analysis of human dendritic cells activated with PPAR-γ ligands showed that PPAR-γ largely affects dendritic-cell development and function through the direct regulation of lipid metabolism, rather than through transrepression[42]. These studies provide considerable support for a growing body of literature indicating that PPAR-γ signals are important factors in inflammation and immunity.

Biology of PPAR-α

The gene encoding PPAR-α was initially cloned from mouse liver cDNA on the basis of its properties as a nuclear receptor activated by carcinogens that induce peroxisome proliferation in the liver[9]. PPAR-α is

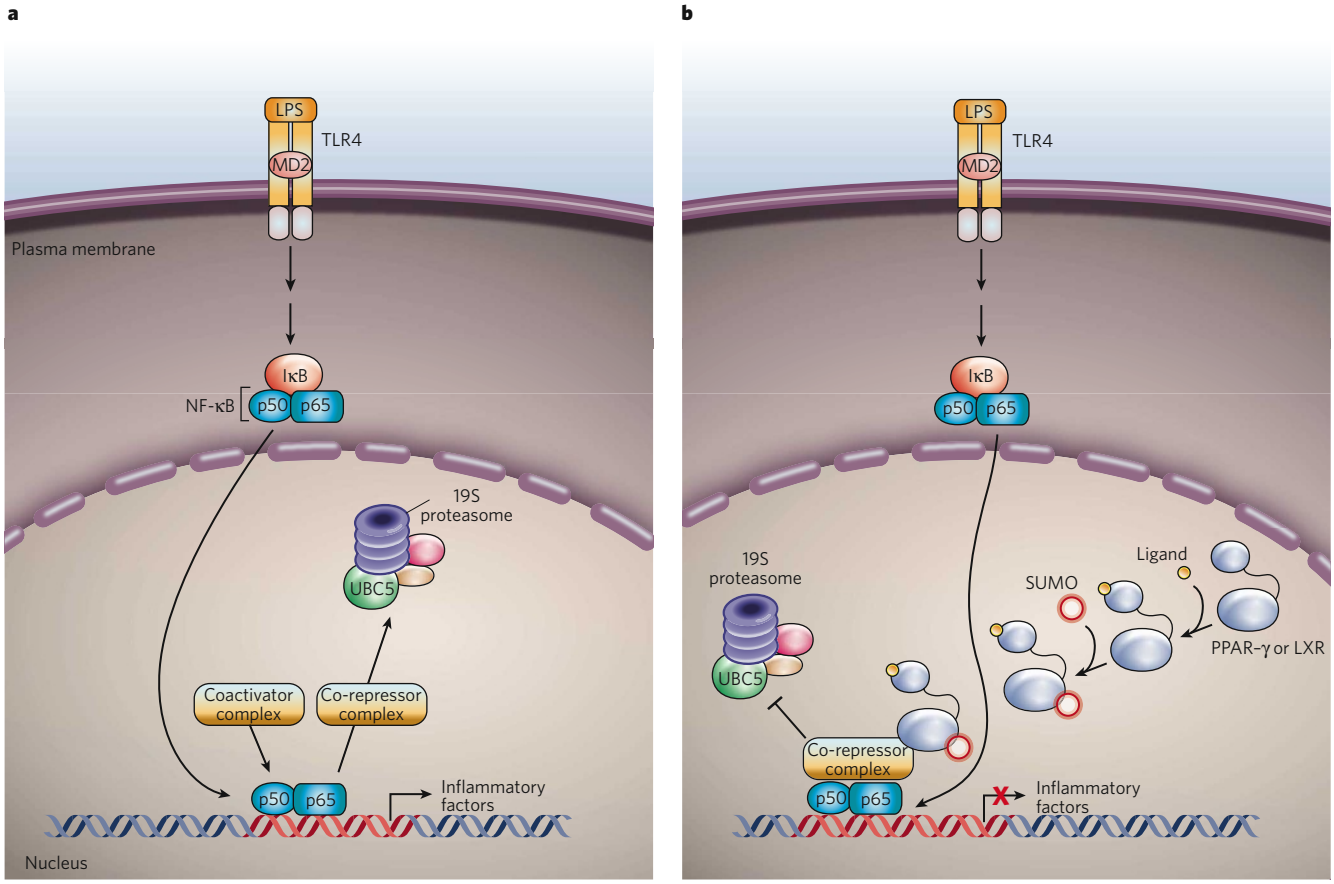


Figure 1 | A model for signal-specific PPAR-γ- and LXR-mediated transrepression. **a**, In the basal state, inflammatory genes are repressed by co-repressor complexes that contain histone deacetylases (HDACs), and the nuclear-receptor co-repressor (NCOR) or silencing mediator of retinoic acid and thyroid-hormone receptor (SMRT). These protein complexes are bound to the promoters of inflammatory genes through interaction with undefined nuclear factors. Inflammatory signalling, for example lipopolysaccharide (LPS)-mediated signalling through Toll-like receptor 4 (TLR4), results in the clearance and degradation of the co-repressor complex, in a ubiquitin-conjugating enzyme 5 (UBC5)-dependent manner by the 19S proteasome. LPS-mediated signalling also triggers the degradation of inhibitor of NF-κB (IκB) and subsequent translocation of NF-κB (p50–p65 heterodimers) to gene promoter sites, as well as the assembly of a coactivator complex on these sites, together resulting in inflammatory gene transcription. **b**, The activation of PPAR-γ inhibits the expression of inflammatory genes described in **a**. According to the model proposed by Glass and colleagues[26], the binding of ligand to PPAR-γ changes the conformation of PPAR-γ such that SUMO can bind. SUMOylated PPAR-γ then binds to the NCOR–HDAC-containing co-repressor complex, preventing its degradation by the 19S proteasome and thereby maintaining the inflammatory genes in a repressed state.

predominantly expressed in the liver, brown fat and heart, and has been implicated in regulating cellular energetics. As such, many PPAR-α target genes are involved in mitochondrial and peroxisomal β-oxidation of fatty acids, including those encoding carnitine palmitoyltransferase 1B and acyl-coenzyme A oxidase[43]. Numerous studies of *Ppara*$^{-/-}$ mice have shown that fasting or high-fat feeding of these mice results in abnormal lipid accumulation in hepatocytes, which is consistent with a crucial role for PPAR-α in hepatic lipid metabolism[44,45]. In addition to the control of cellular fatty-acid metabolism, PPAR-α regulates systemic lipid metabolism by controlling the expression of apolipoprotein and lipoprotein lipase[46,47]. Finally, PPAR-α activation seems to influence apoptosis in macrophages exposed to inflammatory cytokines, including tumour necrosis factor-α (TNF-α) and interferon-γ[48].

PPAR-α is the molecular target of the fibrates, a class of hypolipidaemic drugs used to treat dyslipidaemia in humans[49]. Fibrate administration lowers triglyceride levels, presumably through its effects on fatty-acid and lipoprotein metabolism. Furthermore, clinical studies have shown that fibrates reduce the incidence of cardiovascular events and atherosclerosis. Endogenous PPAR-α ligands include polyunsaturated fatty acids such as linoleic acid, leukotriene derivatives and VLDLs that are in the presence of enzymatically active lipoprotein lipase (see ref. 5 for a review).

PPAR-α in inflammation and macrophages

PPAR-α is expressed in human and mouse immune cells, including lymphocytes, macrophages and dendritic cells. Numerous studies have implicated PPAR-α in the negative regulation of inflammatory responses. *Ppara*$^{-/-}$ mice have an increased susceptibility to chemically induced colitis, experimental autoimmune encephalitis (EAE, a model of multiple sclerosis) and experimentally induced allergic asthma, strongly suggesting that endogenous PPAR-α quells inflammatory signalling in these model systems. Similarly, gain-of-function studies using PPAR-α ligands have shown a reduction in the symptoms of inflammation and disease in a range of models, including allergic airway disease, arthritis, inflammatory bowel disease and EAE (see ref. 27 for a review). Like PPAR-γ, ligand binding to PPAR-α seems to block the expression of inflammatory cytokines such as IL-6 in a signal-specific manner.

Given that PPAR-α regulates numerous genes involved in lipid metabolism in the liver, fat and muscles, the finding that, in mouse models, global loss of PPAR-α increases susceptibility to developing atherosclerosis was not surprising[50]. However, it was unexpected that selective expression or loss of *Ppara* in bone-marrow-derived cells influences the progression of atherosclerosis in susceptible mice[51]. Mechanistic studies on macrophages derived from *Ppara*$^{-/-}$ bone marrow showed a moderate influence of PPAR-α on foam-cell formation and cholesterol efflux; furthermore, the anti-inflammatory effect of PPAR-α in peritoneal macrophages was clearly evident in response to LPS challenge. These studies further implicate PPAR-α as an important regulator of inflammatory disease.

How PPAR-α influences inflammatory gene expression remains incompletely understood. Several mechanisms are known to be involved, including direct interactions with the transcription factors NF-κB and AP1, alterations in cytokine-receptor and growth-factor receptor signalling, and upregulation of expression of a subunit of inhibitor of NF-κB (IκB) (see ref. 6 for a review). SUMOylation-dependent transrepression by PPAR-α has not been shown at this time. However, it is interesting to note that PPAR-α and PPAR-δ have a high degree of sequence homology at sites surrounding the lysine residues involved in SUMOylation-dependent transrepression by PPAR-γ.

Biology of PPAR-δ

PPAR-δ (also known as PPAR-β) was originally identified by Walter Wahli and colleagues in *Xenopus laevis*[52]. The mouse and human receptors were cloned subsequently on the basis of sequence similarity with PPAR-α[10]. PPAR-δ is ubiquitously expressed, suggesting a fundamental requirement for PPAR-δ signalling in many tissues. Consistent with this concept, *Ppard*$^{-/-}$ mice have placental defects that result in increased embryonic lethality and myelination defects,

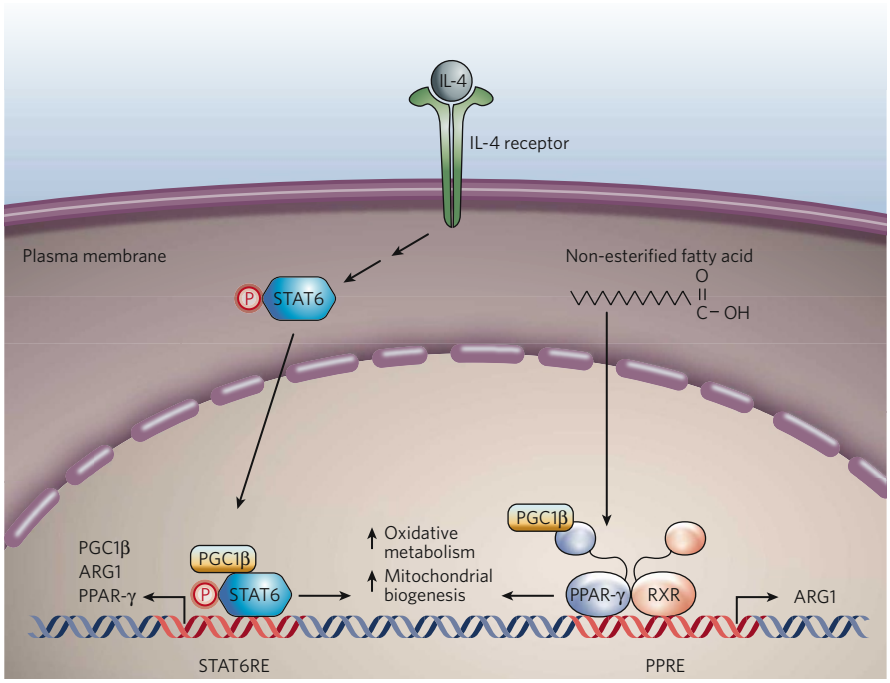


Figure 2 | Integration of lipid and cytokine signals by PPAR-γ and PGC1β. The figure depicts a model for the alternative activation of monocytes to become macrophages of the M2 phenotype. IL-4-mediated signalling activates the transcription factor STAT6, resulting in upregulation of the expression of PPAR-γ, the coactivator PGC1β and ARG1. Increased PGC1β expression reinforces the action of STAT6 on these genes, as well as on other genes involved in mitochondrial biogenesis, oxidative metabolism and alternative activation. At the same time, PPAR-γ is activated by endogenous lipid ligands (such as non-esterified fatty acids), thereby promoting mitochondrial biogenesis, oxidative metabolism and expression of the target genes involved in M2 macrophage function (such as the gene encoding ARG1). Targeted disruption of the genes encoding STAT6, PPAR-γ or PGC1β significantly decreases the efficiency of alternative activation and innate immune responses.

decreased amounts of adipose tissue, and altered wound healing and responses to skin inflammation[53–56]. PPAR-δ target genes in metabolic tissues are broadly involved in fatty-acid metabolism, mitochondrial respiration, thermogenesis and the programming of muscle fibre type. Somewhat surprisingly, *Ppard*$^{-/-}$ mice show no significant alterations in high-density lipoprotein (HDL) cholesterol or triglyceride levels. Similarly to PPAR-α and PPAR-γ, the naturally occurring ligands for PPAR-δ probably include polyunsaturated fatty acids and eicosanoids (Table 1). PPAR-δ has also been shown to be activated by VLDLs in the presence of enzymatically active lipoprotein lipase[57]. Highly potent and selective synthetic PPAR-δ agonists have been developed and used to elucidate PPAR-δ target genes and biology. In animal models of cardiovascular disease or obesity, systemic administration of PPAR-δ ligands implicated PPAR-δ signalling in lipoprotein metabolism. Strikingly, administration of the synthetic ligand GW501516 to rhesus macaques resulted in a significant increase in HDL cholesterol, a decrease in LDL cholesterol and a decrease in fasting triglyceride concentrations[58].

PPAR-δ in inflammation and macrophage biology

In contrast to PPAR-γ, the role of PPAR-δ in the modulation of inflammation is poorly understood. Chih-Hao Lee and colleagues found that loss of haematopoietic PPAR-δ expression protected against atherosclerosis[25]. Analysis of *Ppard*$^{-/-}$ macrophages suggested that the protective effects were not attributable to alterations in lipid metabolism. Rather, the authors proposed that the pro-atherogenic effects of PPAR-δ were due in part to the influence of PPAR-δ on the basal expression of inflammatory mediators in the arterial wall. Interestingly, the authors found reduced expression of CCL2, matrix metalloproteinase 9 (MMP9) and IL-1β in *Ppard*$^{-/-}$ bone-marrow-derived macrophages. Mechanistic studies showed an interaction between PPAR-δ and the transcriptional repressor BCL-6. The authors proposed that the PPAR-δ–BCL-6 interaction sequesters BCL-6 away from the promoters of inflammatory genes. In this system, ligand-driven activation of PPAR-δ releases the co-repressor and allows the repression of inflammatory genes. This model is in contrast to that proposed by Glass and colleagues for PPAR-γ[26], and future studies should focus on testing these working models.

As discussed earlier, monocytes can differentiate into macrophages with classically activated (M1) or alternatively activated (M2) phenotypes, and the balance between these phenotypes seems to depend on the level of PPAR-γ expression[31]. Whether differentiation into M2 macrophages is also influenced by the activity of other PPARs remains an open question. Interestingly, evidence from Inés Corraliza and colleagues suggests that some aspects of alternative activation are indeed PPAR-δ-dependent[59]. In these studies, natural and synthetic PPAR-δ ligands were shown to upregulate expression of arginase 1 (ARG1), a marker of alternative activation of macrophages, in a PPAR-δ-dependent manner. IL-4-mediated upregulation of ARG1 expression was also abrogated by deletion of the gene encoding PPAR-δ or PPAR-γ. Interestingly, sustained PPAR-δ signalling was found to result in reduced clearance of *L. major*, an intracellular parasite that needs to be taken up by M1 macrophage to be cleared. However, PPAR-α ligands did not modulate ARG1 expression, which indicates that PPAR-α does not contribute to this aspect of differentiation into M2 macrophages in mice cells. Taken together, these data suggest that PPAR-δ and PPAR-γ have complementary but not compensatory roles in the acquisition of the alternatively activated macrophages.

LXRs

The genes that encode the nuclear receptors LXR-α and LXR-β (also known as NR1H3 and NR1H2, respectively) were cloned more than a decade ago on the basis of sequence homology with other nuclear receptors. Since then, a large body of work has established that LXRs are cholesterol sensors that regulate both cellular and systemic cholesterol homeostasis. Not surprisingly, the natural ligands for LXR are sterol metabolites such as 22-(*R*)-hydroxycholesterol, 24-(*S*)-hydroxycholesterol, 27-hydroxycholesterol and 24-(*S*),25-epoxycholesterol[60–62]. LXR-α and LXR-β have considerable sequence homology (~77% identity in DNA- and ligand-binding domains), seem to respond to the same endogenous ligands, and have almost identical target genes[34,63,64]. However, an important distinction is their tissue distribution. LXR-β is ubiquitously expressed, whereas LXR-α is restricted to the liver, adipose tissue, adrenal glands, intestine, lungs, kidneys and cells of myeloid origin. The mechanisms that control LXR expression are poorly understood. But it is interesting to note that human LXR-α can autoregulate its expression, whereas human LXR-β and murine LXR-α and LXR-β cannot.

LXRs as regulators of whole-body cholesterol homeostasis

LXRs regulate the expression of genes involved in cholesterol metabolism in a tissue-specific manner. For example, LXR activation in rodent liver upregulates the expression of CYP7A1 (ref. 65) (a member of the cytochrome P450 family that is important for bile-acid synthesis) and the transporters ABCG5 and ABCG8 (which are important for the secretion of cholesterol into bile). In the intestine, LXRs control the reabsorption of cholesterol through the expression of ABCG5 and ABCG8 (refs 66, 67). In peripheral cells such as macrophages, LXRs control the expression of a set of genes involved in the return of peripheral cholesterol to the liver by a process known as reverse cholesterol transport. In response to cholesterol loading, LXRs induce expression of the cholesterol-efflux transporters ABCA1 and ABCG1, the lipoprotein-remodelling enzyme PLTP, and apolipoproteins of the APOC subfamily and APOE[5]. Systemic activation of LXRs thus initiates a series of tissue-specific transcriptional programs that regulate whole-body cholesterol content. Pharmacological activation of these receptors *in vivo* results in increased HDL levels and net cholesterol loss.

Prompted by the compelling biological functions of LXRs, researchers have begun to explore the potential uses of synthetic LXR ligands (Table 1) to treat dyslipidaemia and atherosclerosis. Systemic administration of synthetic LXR agonists protects against atherosclerosis, effectively blocking atherosclerotic lesion development in susceptible mice[68,69]. More importantly, these ligands can also result in the regression of established lesions in mice. However, hepatic activation of LXR also increases plasma triglyceride levels, an independent risk factor for cardiovascular disease. Studies of *Lxra*$^{-/-}$ and *Lxrb*$^{-/-}$ mice have shown that the hepatic lipogenic activity of LXRs is predominantly under the control of LXR-α, which suggests that selective LXR-β agonists could have substantial therapeutic benefit without the undesirable side effects associated with LXR-α activation[65,70]. But is selective LXR-β activation sufficient to ameliorate atherosclerosis? A recent series of studies has begun to address this issue and the closely related question of whether selective inactivation of LXR-α exacerbates atherosclerosis in susceptible mice[71]. Interestingly, ageing *Apoe*$^{-/-}$*Lxra*$^{-/-}$ mice accumulated significantly more lipids in peripheral tissues and developed atherosclerosis faster than *Apoe*$^{-/-}$ mice. These data suggest that LXR-α is essential for optimal reverse cholesterol transport. Perhaps more importantly, systemic administration of an LXR ligand decreased peripheral cholesterol content and had anti-atherogenic effects without significant increasing plasma triglyceride levels[71]. Thus, the lipogenic activity of LXRs can be uncoupled from the cholesterol-homeostasis program. These studies provide an important proof of principle for the therapeutic use of selective LXR-β agonists.

LXRs in inflammation and transrepression

In addition to modulating cholesterol homeostasis, LXRs have emerged as important regulators of inflammatory gene expression and innate immunity. We and others have shown that activation of LXRs antagonizes inflammatory gene expression downstream of TLR4 signalling, IL-1β-mediated signalling and TNF-α-mediated signalling[72,73]. Inflammatory genes repressed by LXRs include several NF-κB target genes, such as iNOS, IL-6, cyclooxygenase 1, MMP9, CCL2, CCL7 and IL-1β. Conversely, activation of either TLR3 and TLR4 inhibits the function of LXRs in cholesterol homeostasis through the transcription factor interferon-regulatory factor 3, suggesting that LXRs regulate cross-talk between inflammatory and metabolic pathways. Interestingly, LXRs do not seem to influence inflammatory gene expression downstream of

TLR3, indicating that transrepression by LXRs is context specific. Both loss-of-function and gain-of-function studies have established a role for LXR signalling in modulating inflammation in models of contact dermatitis and atherosclerosis[68,72,74].

As discussed earlier, the mechanism of PPAR-γ-mediated transrepression depends on protein–protein interactions that preserve the integrity of the co-repressor complex while it is associated with the promoters of inflammatory genes. A similar, but not identical, mechanism has recently been elucidated for LXRs[75]. Ligand binding to LXRs results in SUMOylation by SUMO2 or SUMO3 and depends on the E3-ligase activity of HDAC4 (rather than on SUMO1 and PIAS1, as is the case for PPAR-γ) for the preservation of the co-repressor complex during inflammatory signalling. Although PPAR-γ and LXRs can transrepress several inflammatory genes in a similar manner, comparative DNA-microarray studies have identified overlapping, but distinct, subsets of genes that are repressed by ligand binding. The specific molecular requirements for transrepression by PPAR-γ and LXRs help to explain this observation. Why these nuclear receptors use parallel molecular mechanisms to negatively regulate similar but distinct gene subsets in the same cell type remains an open and intriguing question. Nevertheless, these data suggest that LXR and PPAR-γ regulate inflammation in different ways and in response to distinct signalling pathways.

LXRs and neuropathology

Recent studies of other inflammatory conditions, such as Alzheimer's disease, have highlighted an important role for LXR signalling in controlling inflammation in the brain. Alzheimer's disease is an age-dependent neurodegenerative disorder that results in progressive cognitive deficits[76]. The pathology of Alzheimer's disease is characterized by extraneuronal deposition of amyloid-β fibrils and intraneuronal tangles of hyperphosphorylated tau[77]. Alzheimer's disease is a multifactorial disease, and both cholesterol metabolism and inflammation seem to contribute to its pathogenesis.

Early studies *in vitro* indicated that LXRs might be involved in disease progression. LXR activation by endogenous ligands or synthetic ligands was reported to upregulate the cholesterol-efflux transporter ABCA1 and to decrease amyloid-β secretion from primary neural-cell cultures[78,79]. In complementary studies, culturing primary astrocytes and microglia with a synthetic LXR ligand reduced inflammatory gene products in response to TLR4 signals[80]. Systemic administration of an LXR agonist led to similar upregulation of expression of the cholesterol-efflux pathway genes and a decrease in amyloid-β secretion in the *App23*-transgenic mouse model of Alzheimer's disease[81]. More recently, we reported that deletion of either *Lxra* or *Lxrb* exacerbates Alzheimer's disease pathology[82]. An examination of microglia and astrocytes, key components of the neural innate immune system, showed that LXRs were present at levels comparable to those of peripheral macrophages. More importantly, LXR activation attenuated the expression of inflammatory gene products in response to LPS and to fibrillar amyloid-β peptide (a small synthetic peptide derived from the amyloid-β protein that is designed to mimic the activity of the whole protein) in microglia. Interestingly, LXR activation seems to preserve the phagocytic capacity of microglia in an otherwise inflammatory environment. How this occurs mechanistically remains an open question. Nevertheless, it seems to be an important component of microglial biology.

In another series of studies, systemic administration of LXR ligands was found to decrease the severity of EAE[83]. In these studies, the authors noted a reduction in demyelination, inflammation of the central nervous system and expression of major histocompatibility complex (MHC) class II molecules by microglia, and an attenuation of clinical disease. Thus, it seems that LXRs play a substantial part in attenuating neuroinflammation and might be therapeutic targets for reducing inflammatory pathology in the CNS.

LXRs and host defence

The ability of LXRs to negatively regulate inflammatory gene expression and limit neural immunopathology suggests that LXR activation might be deleterious to host defence. Thus, it was surprising to find that loss of LXR function compromised innate immunity. Mice lacking LXRs are more susceptible to challenge with the Gram-positive intracellular pathogen *Listeria monocytogenes*[84]. Reciprocal bone-marrow transplant studies (wild-type bone marrow into $Lxra^{-/-}Lxrb^{-/-}$ mice and $Lxra^{-/-}Lxrb^{-/-}$ bone marrow into wild-type mice) showed that protection against *L. monocytogenes* challenge requires LXR expression by haematopoietic cells. In particular, LXR-α seems to provide the greatest protection, and loss of LXR-α correlates with increased macrophage apoptosis. The increased susceptibility of $Lxra^{-/-}Lxrb^{-/-}$ macrophages to pathogen-induced apoptosis results, at least in part, from the loss of regulation of the anti-apoptotic gene *Aim* by LXR-α. Similar studies by Annabel Valledor *et al.*[85] showed that LXR signalling also inhibits macrophage apoptosis in response to drug-induced apoptotic stimuli (for example, cycloheximide), cytokine withdrawal, and infection with the bacteria *Bacillus anthracis*, *Escherichia coli* or *Salmonella enterica* serovar Typhimurium. This activity was attributed to the induction of anti-apoptotic genes (including *Aim*) and the inhibition of a subset of pro-apoptotic genes.

The effects of LXR signalling on inflammation and host defence have been further highlighted by two recent studies on pulmonary inflammation[86,87]. In both studies, systemic administration of LXR agonists reduced localized lung inflammatory responses to LPS exposure. The reduction in inflammation was mediated in part by an expected reduction in inflammatory gene expression. However, it was intriguing to note that both studies found a decrease in airway neutrophilia in the LXR-ligand-treated groups. Kathleen Smoak *et al.* showed also an attenuation of neutrophil migration in response to *Klebsiella pneumoniae* challenge, resulting in impaired host defence[87]. LXRs also seem to regulate the maturation and function of human myeloid dendritic cells. Synthetic LXR agonists have been reported to decrease LPS-induced IL-12 secretion, increase IL-10 secretion and alter T-cell activation[88]. Thus, it seems that LXR activation may either impair or improve host defence, depending on the context.

Conclusions

A growing body of literature supports the idea that lipid metabolism and inflammation are closely linked and that cross-talk between these processes is fundamental to the pathogenesis of human diseases such as type 2 diabetes and atherosclerosis. Conceptually, it is not entirely clear why metabolism and inflammation should be so tightly linked by the nuclear receptors of the LXR and PPAR families, particularly in macrophage biology. One possibility is that the metabolic and inflammatory signalling functions of nuclear receptors have evolved independently. However, we favour an alternative hypothesis: that a cell's underlying metabolic state is vital for efficient cellular function. In this way, the PPAR and LXR nuclear receptors function as transducers and reinforcers of particular metabolic programs required to support particular inflammatory programs. In addition, these nuclear receptors may modulate inflammatory signals that alter or disrupt a cell's 'preferred' metabolic state. Studies of macrophages[31,36] have supported these ideas, and it will be important to test these ideas in other cell types.

Several intriguing issues regarding LXR and PPAR signalling remain to be addressed. These include whether selective LXR-β agonists can be developed to treat individuals with atherosclerosis and whether the clinically available PPAR-γ agonist drugs can be used to treat individuals with inflammatory disorders. Another question is whether synthetic ligands that uncouple the anti-inflammatory effects of LXRs from their role in cholesterol homeostasis can be developed. It will also be important to further elucidate the complex mechanism of transrepression. Elegant studies have provided an important working model for PPAR- and LXR-mediated transrepression. However, the components of the machinery remain to be fully elucidated, and it will be important to test this model in the context of normal biology, disease and therapy. Finally, the relationship between PPARs and macrophage differentiation remains to be unravelled. For example, is PPAR-γ signalling a requirement for the alternative activation of macrophages, or is it permissive only in certain contexts? Does the loss of *Ppard* result in the same phenotype as the loss of *Pparg* in these

model systems? Preliminary data from studies of infection[59] suggest that both PPAR-γ and PPAR-δ are important for immune-cell function, and it will be interesting to determine whether selective loss of *Pparg* in the haematopoietic system also affects adiposity or metabolic syndrome. If so, are these receptors transcriptionally regulating genes in the same pathways or in parallel differentiation pathways? Further studies will be required to fully understand these and other issues regarding the complex cellular and molecular biology of these nuclear receptors in both normal conditions and disease states. ■

1. Evans, R. M. The steroid and thyroid hormone receptor superfamily. *Science* **240,** 889–895 (1988).
2. Mangelsdorf, D. J. & Evans, R. M. The RXR heterodimers and orphan receptors. *Cell* **83,** 841–850 (1995).
3. Mangelsdorf, D. J. *et al.* The nuclear receptor superfamily: the second decade. *Cell* **83,** 835–839 (1995).
4. Giguere, V. Orphan nuclear receptors: from gene to function. *Endocr. Rev.* **20,** 689–725 (1999).
5. Castrillo, A. & Tontonoz, P. Nuclear receptors in macrophage biology: at the crossroads of lipid metabolism and inflammation. *Annu. Rev. Cell Dev. Biol.* **20,** 455–480 (2004).
6. Glass, C. K. & Ogawa, S. Combinatorial roles of nuclear receptors in inflammation and immunity. *Nature Rev. Immunol.* **6,** 44–55 (2006).
7. Glass, C. K. & Rosenfeld, M. G. The coregulator exchange in transcriptional functions of nuclear receptors. *Genes Dev.* **14,** 121–141 (2000).
8. Berger, J. & Moller, D. E. The mechanisms of action of PPARs. *Annu. Rev. Med.* **53,** 409–435 (2002).
9. Issemann, I. & Green, S. Activation of a member of the steroid hormone receptor superfamily by peroxisome proliferators. *Nature* **347,** 645–650 (1990).
10. Kliewer, S. A. *et al.* Differential expression and activation of a family of murine peroxisome proliferator-activated receptors. *Proc. Natl Acad. Sci. USA* **91,** 7355–7359 (1994).
11. Tontonoz, P., Hu, E., Graves, R. A., Budavari, A. I. & Spiegelman, B. M. mPPARγ2: tissue-specific regulator of an adipocyte enhancer. *Genes Dev.* **8,** 1224–1234 (1994).
12. Kliewer, S. A., Umesono, K., Noonan, D. J., Heyman, R. A. & Evans, R. M. Convergence of 9-*cis* retinoic acid and peroxisome proliferator signalling pathways through heterodimer formation of their receptors. *Nature* **358,** 771–774 (1992).
13. Rosen, E. D. & Spiegelman, B. M. Molecular regulation of adipogenesis. *Annu. Rev. Cell Dev. Biol.* **16,** 145–171 (2000).
14. Willson, T. M., Lambert, M. H. & Kliewer, S. A. Peroxisome proliferator-activated receptor γ and metabolic disease. *Annu. Rev. Biochem.* **70,** 341–367 (2001).
15. Tontonoz, P., Hu, E. & Spiegelman, B. M. Stimulation of adipogenesis in fibroblasts by PPARγ2, a lipid-activated transcription factor. *Cell* **79,** 1147–1156 (1994).
16. Barak, Y. *et al.* PPARγ is required for placental, cardiac, and adipose tissue development. *Mol. Cell* **4,** 585–595 (1999).
17. Lehmann, J. M. *et al.* An antidiabetic thiazolidinedione is a high affinity ligand for peroxisome proliferator-activated receptor γ (PPARγ). *J. Biol. Chem.* **270,** 12953–12956 (1995).
18. Guan, H. P., Ishizuka, T., Chui, P. C., Lehrke, M. & Lazar, M. A. Corepressors selectively control the transcriptional activity of PPARγ in adipocytes. *Genes Dev.* **19,** 453–461 (2005).
19. Delerive, P. *et al.* Peroxisome proliferator-activated receptor α negatively regulates the vascular inflammatory gene response by negative cross-talk with transcription factors NF-κB and AP-1. *J. Biol. Chem.* **274,** 32048–32054 (1999).
20. Chung, S. W. *et al.* Oxidized low density lipoprotein inhibits interleukin-12 production in lipopolysaccharide-activated mouse macrophages via direct interactions between peroxisome proliferator-activated receptor-γ and nuclear factor-κB. *J. Biol. Chem.* **275,** 32681–32687 (2000).
21. Zingarelli, B. *et al.* Peroxisome proliferator activator receptor-γ ligands, 15-deoxy-$\Delta^{12,14}$-prostaglandin J_2 and ciglitazone, reduce systemic inflammation in polymicrobial sepsis by modulation of signal transduction pathways. *J. Immunol.* **171,** 6827–6837 (2003).
22. Kelly, D. *et al.* Commensal anaerobic gut bacteria attenuate inflammation by regulating nuclear-cytoplasmic shuttling of PPAR-γ and RelA. *Nature Immunol.* **5,** 104–112 (2004).
23. Syrovets, T., Schule, A., Jendrach, M., Buchele, B. & Simmet, T. Ciglitazone inhibits plasmin-induced proinflammatory monocyte activation via modulation of p38 MAP kinase activity. *Thromb. Haemost.* **88,** 274–281 (2002).
24. Li, M., Pascual, G. & Glass, C. K. Peroxisome proliferator-activated receptor γ-dependent repression of the inducible nitric oxide synthase gene. *Mol. Cell Biol.* **20,** 4699–4707 (2000).
25. Lee, C.-H. *et al.* Transcriptional repression of atherogenic inflammation: modulation by PPARδ. *Science* **302,** 453–457 (2003).
This paper describes how unliganded PPAR-δ is inflammatory: PPAR-δ sequesters the transcriptional repressor BCL-6 away from the promoters of inflammatory genes. Ligand binding to PPAR-δ releases BCL-6, resulting in the repression of inflammatory gene expression.
26. Pascual, G. *et al.* A SUMOylation-dependent pathway mediates transrepression of inflammatory response genes by PPAR-γ. *Nature* **437,** 759–763 (2005).
This paper shows that ligand-driven PPAR-γ transrepression of inflammatory genes occurs by a SUMOylation- and NCOR-dependent pathway.
27. Straus, D. S. & Glass, C. K. Anti-inflammatory actions of PPAR ligands: new insights on cellular and molecular mechanisms. *Trends Immunol.* **28,** 551–558 (2007).
28. Adachi, M. *et al.* Peroxisome proliferator activated receptor γ in colonic epithelial cells protects against experimental inflammatory bowel disease. *Gut* **55,** 1104–1113 (2006).
29. Shah, Y. M., Morimura, K. & Gonzalez, F. J. Expression of peroxisome proliferator-activated receptor-γ in macrophage suppresses experimentally induced colitis. *Am. J. Physiol. Gastrointest. Liver Physiol.* **292,** G657–G666 (2007).
30. Wan, Y. *et al.* Maternal PPARγ protects nursing neonates by suppressing the production of inflammatory milk. *Genes Dev.* **21,** 1895–1908 (2007).
This paper shows that endothelial-specific deletion of the gene encoding PPAR-γ modulates the presence of inflammatory lipids in milk produced by mammary glands.
31. Odegaard, J. I. *et al.* Macrophage-specific PPARγ controls alternative activation and improves insulin resistance. *Nature* **447,** 1116–1120 (2007).
This paper reports that the IL-4–STAT6–PPAR-γ signalling axis in monocytes is crucial for their differentiation into alternatively activated macrophages and for innate immunity. It also shows that PPAR-γ signalling in macrophages modulates diet-induced obesity and peripheral insulin resistance.
32. Huang, J. T. *et al.* Interleukin-4-dependent production of PPAR-γ ligands in macrophages by 12/15-lipoxygenase. *Nature* **400,** 378–382 (1999).
This paper shows that IL-4-mediated signalling upregulates PPAR-γ expression, and it provides the initial evidence that the IL-4–PPAR-γ signalling pathway regulates macrophage function.
33. Shoelson, S. E., Lee, J. & Goldfine, A. B. Inflammation and insulin resistance. *J. Clin. Invest.* **116,** 1793–1801 (2006).
34. Zelcer, N. & Tontonoz, P. Liver X receptors as integrators of metabolic and inflammatory signaling. *J. Clin. Invest.* **116,** 607–614 (2006).
35. Gordon, S. Alternative activation of macrophages. *Nature Rev. Immunol.* **3,** 23–35 (2003).
36. Vats, D. *et al.* Oxidative metabolism and PGC-1β attenuate macrophage-mediated inflammation. *Cell Metab.* **4,** 13–24 (2006).
37. Bouhlel, M. A. *et al.* PPARγ activation primes human monocytes into alternative M2 macrophages with anti-inflammatory properties. *Cell Metab.* **6,** 137–143 (2007).
38. Faveeuw, C. *et al.* Peroxisome proliferator-activated receptor γ activators inhibit interleukin-12 production in murine dendritic cells. *FEBS Lett.* **486,** 261–266 (2000).
39. Gosset, P. *et al.* Peroxisome proliferator-activated receptorγ activators affect the maturation of human monocyte-derived dendritic cells. *Eur. J. Immunol.* **31,** 2857–2865 (2001).
40. Szatmari, I. *et al.* Activation of PPARγ specifies a dendritic cell subtype capable of enhanced induction of iNKT cell expansion. *Immunity* **21,** 95–106 (2004).
41. Klotz, L. *et al.* Peroxisome proliferator-activated receptor γ control of dendritic cell function contributes to development of $CD4^+$ T cell anergy. *J. Immunol.* **178,** 2122–2131 (2007).
42. Szatmari, I. *et al.* PPARγ regulates the function of human dendritic cells primarily by altering lipid metabolism. *Blood* **110,** 3271–3280 (2007).
43. Dreyer, C. *et al.* Positive regulation of the peroxisomal β-oxidation pathway by fatty acids through activation of peroxisome proliferator-activated receptors (PPAR). *Biol. Cell* **77,** 67–76 (1993).
44. Kersten, S. *et al.* Peroxisome proliferator-activated receptor α mediates the adaptive response to fasting. *J. Clin. Invest.* **103,** 1489–1498 (1999).
45. Leone, T. C., Weinheimer, C. J. & Kelly, D. P. A critical role for the peroxisome proliferator-activated receptor α (PPARα) in the cellular fasting response: the PPARα-null mouse as a model of fatty acid oxidation disorders. *Proc. Natl Acad. Sci. USA* **96,** 7473–7478 (1999).
46. Staels, B., van Tol, A., Andreu, T. & Auwerx, J. Fibrates influence the expression of genes involved in lipoprotein metabolism in a tissue-selective manner in the rat. *Arterioscler. Thromb.* **12,** 286–294 (1992).
47. Schoonjans, K. *et al.* PPARα and PPARγ activators direct a distinct tissue-specific transcriptional response via a PPRE in the lipoprotein lipase gene. *EMBO J.* **15,** 5336–5348 (1996).
48. Chinetti, G. *et al.* Activation of proliferator-activated receptors α and γ induces apoptosis of human monocyte-derived macrophages. *J. Biol. Chem.* **273,** 25573–25580 (1998).
49. Forman, B. M., Chen, J. & Evans, R. M. Hypolipidemic drugs, polyunsaturated fatty acids, and eicosanoids are ligands for peroxisome proliferator-activated receptors α and δ. *Proc. Natl Acad. Sci. USA* **94,** 4312–4317 (1997).
50. Tordjman, K. *et al.* PPARα deficiency reduces insulin resistance and atherosclerosis in apoE-null mice. *J. Clin. Invest.* **107,** 1025–1034 (2001).
51. Babaev, V. R. *et al.* Macrophage expression of peroxisome proliferator-activated receptor-α reduces atherosclerosis in low-density lipoprotein receptor-deficient mice. *Circulation* **116,** 1404–1412 (2007).
52. Dreyer, C. *et al.* Control of the peroxisomal β-oxidation pathway by a novel family of nuclear hormone receptors. *Cell* **68,** 879–887 (1992).
53. Peters, J. M. *et al.* Growth, adipose, brain, and skin alterations resulting from targeted disruption of the mouse peroxisome proliferator-activated receptor β (δ). *Mol. Cell Biol.* **20,** 5119–5128 (2000).
54. Michalik, L. *et al.* Impaired skin wound healing in peroxisome proliferator-activated receptor (PPAR)α and PPARβ mutant mice. *J. Cell Biol.* **154,** 799–814 (2001).
55. Tan, N. S. *et al.* Critical roles of PPAR β/δ in keratinocyte response to inflammation. *Genes Dev.* **15,** 3263–3277 (2001).
56. Barak, Y. *et al.* Effects of peroxisome proliferator-activated receptor δ on placentation, adiposity, and colorectal cancer. *Proc. Natl Acad. Sci. USA* **99,** 303–308 (2002).
57. Chawla, A. *et al.* PPARδ is a very low-density lipoprotein sensor in macrophages. *Proc. Natl Acad. Sci. USA* **100,** 1268–1273 (2003).
58. Oliver, W. R. Jr *et al.* A selective peroxisome proliferator-activated receptor δ agonist promotes reverse cholesterol transport. *Proc. Natl Acad. Sci. USA* **98,** 5306–5311 (2001).
59. Gallardo-Soler, A. *et al.* Arginase I induction by modified lipoproteins in macrophages: a PPAR-γ/δ-mediated effect that links lipid metabolism and immunity. *Mol. Endocrinol.* **22,** 1394–1402 (2008).
60. Janowski, B. A., Willy, P. J., Devi, T. R., Falck, J. R. & Mangelsdorf, D. J. An oxysterol signalling pathway mediated by the nuclear receptor LXRα. *Nature* **383,** 728–731 (1996).
61. Lehmann, J. M. *et al.* Activation of the nuclear receptor LXR by oxysterols defines a new hormone response pathway. *J. Biol. Chem.* **272,** 3137–3140 (1997).
62. Fu, X. *et al.* 27-hydroxycholesterol is an endogenous ligand for liver X receptor in cholesterol-loaded cells. *J. Biol. Chem.* **276,** 38378–38387 (2001).
63. Repa, J. J. & Mangelsdorf, D. J. The role of orphan nuclear receptors in the regulation of cholesterol homeostasis. *Annu. Rev. Cell Dev. Biol.* **16,** 459–481 (2000).
64. Tontonoz, P. & Mangelsdorf, D. J. Liver X receptor signaling pathways in cardiovascular disease. *Mol. Endocrinol.* **17,** 985–993 (2003).
65. Peet, D. J. *et al.* Cholesterol and bile acid metabolism are impaired in mice lacking the nuclear oxysterol receptor LXRα. *Cell* **93,** 693–704 (1998).

66. Repa, J. J. *et al.* Regulation of absorption and ABC1-mediated efflux of cholesterol by RXR heterodimers. *Science* **289,** 1524–1529 (2000).
67. Repa, J. J. *et al*. Regulation of ATP-binding cassette sterol transporters ABCG5 and ABCG8 by the liver X receptors α and β. *J. Biol. Chem.* **277,** 18793–18800 (2002).
68. Joseph, S. B. *et al.* Synthetic LXR ligand inhibits the development of atherosclerosis in mice. *Proc. Natl Acad. Sci. USA* **99,** 7604–7609 (2002).
69. Terasaka, N. *et al.* T-0901317, a synthetic liver X receptor ligand, inhibits development of atherosclerosis in LDL receptor-deficient mice. *FEBS Lett.* **536,** 6–11 (2003).
70. Alberti, S. *et al*. Hepatic cholesterol metabolism and resistance to dietary cholesterol in LXRβ-deficient mice. *J. Clin. Invest.* **107,** 565–573 (2001).
71. Bradley, M. N. *et al.* Ligand activation of LXRβ reverses atherosclerosis and cellular cholesterol overload in mice lacking LXRα and apoE. *J. Clin. Invest.* **117,** 2337–2346 (2007).
72. Joseph, S. B., Castrillo, A., Laffitte, B. A., Mangelsdorf, D. J. & Tontonoz, P. Reciprocal regulation of inflammation and lipid metabolism by liver X receptors. *Nature Med.* **9,** 213–219 (2003).
73. Ogawa, S. *et al.* Molecular determinants of crosstalk between nuclear receptors and Toll-like receptors. *Cell* **122,** 707–721 (2005).
74. Tangirala, R. K. *et al.* Identification of macrophage liver X receptors as inhibitors of atherosclerosis. *Proc. Natl Acad. Sci. USA* **99,** 11896–11901 (2002).
75. Ghisletti, S. *et al.* Parallel SUMOylation-dependent pathways mediate gene- and signal-specific transrepression by LXRs and PPARγ. *Mol. Cell* **25,** 57–70 (2007).
76. Selkoe, D. J. The molecular pathology of Alzheimer's disease. *Neuron* **6,** 487–498 (1991).
77. Hardy, J. & Selkoe, D. J. The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. *Science* **297,** 353–356 (2002).
78. Koldamova, R. P. *et al.* 22*R*-hydroxycholesterol and 9-*cis*-retinoic acid induce ATP-binding cassette transporter A1 expression and cholesterol efflux in brain cells and decrease amyloid β secretion. *J. Biol. Chem.* **278,** 13244–13256 (2003).
79. Sun, Y., Yao, J., Kim, T. W. & Tall, A. R. Expression of liver X receptor target genes decreases cellular amyloid β peptide secretion. *J. Biol. Chem.* **278,** 27688–27694 (2003).
80. Zhang-Gandhi, C. X. & Drew, P. D. Liver X receptor and retinoid X receptor agonists inhibit inflammatory responses of microglia and astrocytes. *J. Neuroimmunol.* **183,** 50–59 (2007).
81. Koldamova, R. P. *et al.* The liver X receptor ligand T0901317 decreases amyloid β production in vitro and in a mouse model of Alzheimer's disease. *J. Biol. Chem.* **280,** 4079–4088 (2005).
82. Zelcer, N. *et al.* Attenuation of neuroinflammation and Alzheimer's disease pathology by liver X receptors. *Proc. Natl Acad. Sci. USA* **104,** 10601–10606 (2007).
This paper shows that deletion of the genes encoding LXRs exacerbates Alzheimer's disease pathology. Conversely, activation of LXRs in microglial cells attenuates amyloid-β-peptide-driven inflammation and preserves phagocytic function.
83. Hindinger, C. *et al.* Liver X receptor activation decreases the severity of experimental autoimmune encephalomyelitis. *J. Neurosci. Res.* **84,** 1225–1234 (2006).
84. Joseph, S. B. *et al*. LXR-dependent gene expression is important for macrophage survival and the innate immune response. *Cell* **119,** 299–309 (2004).
This paper was the first to report that mice lacking LXRs are more susceptible to challenge with *L. monocytogenes*.
85. Valledor, A. F. *et al*. Activation of liver X receptors and retinoid X receptors prevents bacterial-induced macrophage apoptosis. *Proc. Natl Acad. Sci. USA* **101,** 17813–17818 (2004).
86. Birrell, M. A. *et al.* Novel role for the liver X nuclear receptor in the suppression of lung inflammatory responses. *J. Biol. Chem.* **282,** 31882–31890 (2007).
87. Smoak, K. *et al*. Effects of liver X receptor agonist treatment on pulmonary inflammation and host defense. *J. Immunol.* **180,** 3305–3312 (2008).
88. Geyeregger, R. *et al*. Liver X receptors regulate dendritic cell phenotype and function through blocked induction of the actin-bundling protein fascin. *Blood* **109,** 4288–4295 (2007).

Acknowledgements P.T. is an investigator of the Howard Hughes Medical Institute. Work in the authors' laboratories was supported by National Institutes of Health grants HL66088 and HL30568 (P.T.) and RR021975 (S.J.B.).

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence should be addressed to P.T. (ptontonoz@mednet.ucla.edu).

ARTICLES

High-resolution mapping of meiotic crossovers and non-crossovers in yeast

Eugenio Mancera[1]*, Richard Bourgon[2]*, Alessandro Brozzi[2], Wolfgang Huber[2] & Lars M. Steinmetz[1]

Meiotic recombination has a central role in the evolution of sexually reproducing organisms. The two recombination outcomes, crossover and non-crossover, increase genetic diversity, but have the potential to homogenize alleles by gene conversion. Whereas crossover rates vary considerably across the genome, non-crossovers and gene conversions have only been identified in a handful of loci. To examine recombination genome wide and at high spatial resolution, we generated maps of crossovers, crossover-associated gene conversion and non-crossover gene conversion using dense genetic marker data collected from all four products of fifty-six yeast (*Saccharomyces cerevisiae*) meioses. Our maps reveal differences in the distributions of crossovers and non-crossovers, showing more regions where either crossovers or non-crossovers are favoured than expected by chance. Furthermore, we detect evidence for interference between crossovers and non-crossovers, a phenomenon previously only known to occur between crossovers. Up to 1% of the genome of each meiotic product is subject to gene conversion in a single meiosis, with detectable bias towards GC nucleotides. To our knowledge the maps represent the first high-resolution, genome-wide characterization of the multiple outcomes of recombination in any organism. In addition, because non-crossover hotspots create holes of reduced linkage within haplotype blocks, our results stress the need to incorporate non-crossovers into genetic linkage analysis.

In most eukaryotes, homologous chromosomes exchange genetic information through recombination during meiosis. This process increases genetic diversity by breaking haplotypes, but it may also homogenize alleles through gene conversion[1,2]. Furthermore, recombination is fundamental to sexual reproduction because it provides physical connections between homologues during the first meiotic division, contributing to correct chromosome segregation[3]. In the current model, meiotic recombination starts with the formation of a double-strand break (DSB)[4,5]. The break is then repaired through a series of steps, involving resection, synthesis and ligation, using the homologous chromosome as a template. Repair results in either a crossover—reciprocal exchange accompanied by a tract subject to gene conversion—or a non-crossover—a tract subject to conversion but not associated with reciprocal exchange[4,6]. At least two pathways form crossovers: the Msh4/Msh5-dependent pathway, which proceeds through a double Holliday junction, and the Mus81/Mms4-dependent pathway[7,8]. In contrast, non-crossovers are thought to be the result of synthesis-dependent strand annealing[9]. It is known that DSB[10–14] and crossover rates[15] vary along chromosomes. Non-crossovers and crossover-associated gene conversions have not been characterized genome wide, however, because this requires monitoring recombination between closely spaced markers along the genomes of all four meiotic products[2].

High-resolution mapping of recombination

In *Saccharomyces cerevisiae* we achieved a detailed characterization of recombination outcomes by genotyping ~52,000 markers in all four viable spores derived from 51 meioses of an S288c/YJM789 hybrid strain[16,17] (Fig. 1). Genomic DNA from parental strains and each of the 204 spores was hybridized to high-density microarrays that tile the genomes of both S288c and YJM789 with a median probe offset of 4 base pairs (bp). To infer genotypes from the hybridization intensities of the probes covering each marker (eight probes per marker on average), we developed a new algorithm, ssGenotyping, based on semi-supervised clustering (see Methods). The high density of polymorphism and probes resulted in spore genotypes with a median distance of 78 bp between consecutive markers (Supplementary Fig. 1). This resolution is over 20 times higher than in the current yeast genetic map[15] and more than 360 times higher than in the most recent human crossover map[18].

Owing to their high resolution, our maps invert the traditional relationship between markers and recombination events: there are multiple markers within most recombination events rather than vice versa. This allows characterization of both crossover-associated and non-crossover gene conversion tracts, which are typically thought to be only 1–2-kilobases (kb) long[2]. Genotype calls from all four spores in each wild-type tetrad were used to infer a total of 4,163 crossovers and 2,126 non-crossovers (see Methods). We expect to have detected nearly all crossovers but, because non-crossovers have no effect on flanking markers, to have missed non-crossovers that completely fell between two markers, or non-crossovers in which mismatch repair restored the original genotype. We observed an average of 90.5 crossovers and 46.2 non-crossovers per meiosis. A total of 30.1% of observed crossovers occurred between two consecutive markers, and therefore had no detectable conversion tract. Taking this percentage as an estimate of the fraction of unobserved non-crossovers, we obtained a corrected total, 90.5 crossovers plus 66.1 non-crossovers, which is remarkably similar to a recent estimate of 140–170 DSBs per meiosis[13].

All chromosomes but one had at least one crossover, in agreement with the essential role that crossovers have in chromosome segregation[3]. The average number of crossovers was linearly related to chromosome length, with an intercept of 1.0, corresponding to one obligate crossover, plus an additional 6.1 crossovers per megabase

[1]European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [2]European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge CB10 1SD, UK.
*These authors contributed equally to this work.

(Mb) (Supplementary Fig. 5). Notably, non-crossovers behaved similarly (3.4 non-crossovers per Mb), but with a lower intercept (0.3).

The median size of conversion tracts was 2.0 kb for those associated with crossovers, and 1.8 kb for non-crossover conversion tracts (see Methods). The difference in medians is statistically significant (Wilcoxon rank-sum test, $P < 0.0001$). These sizes are consistent with previous estimates made at a single yeast hotspot[19], but are considerably larger than single-locus estimates in human[20]. Our finding that crossover tracts tend to be larger than non-crossover tracts also corroborates previous, single-locus observations in yeast and human[20,21].

We observed 57 non-crossover conversion tracts larger than 5 kb in size, the largest being 40.8 kb (minimal length). Three of these were

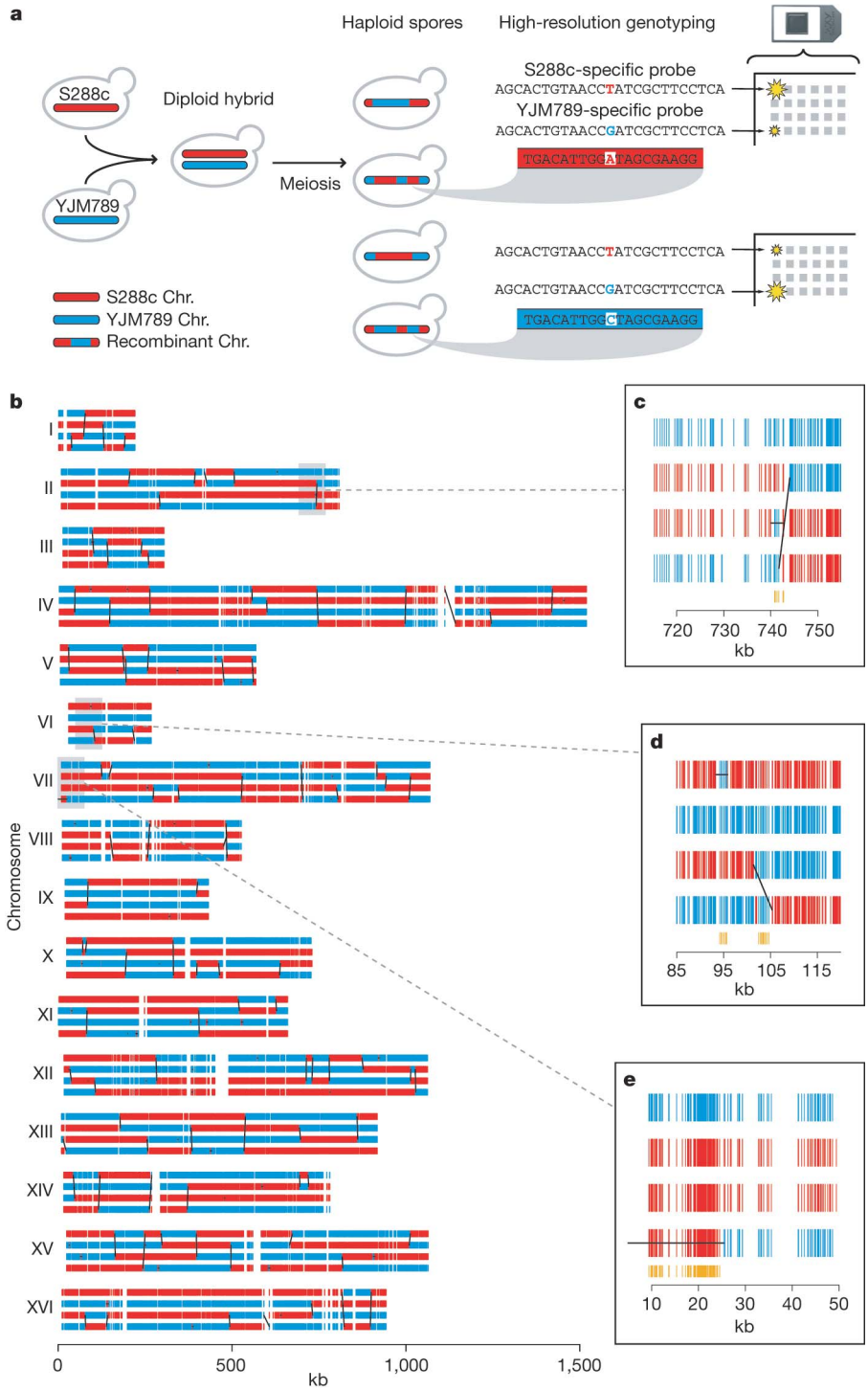


Figure 1 | High-resolution mapping of meiotic recombination along the yeast genome. **a**, Schematic description of the recombination mapping approach. Meiotic products from a hybrid derived from highly polymorphic strains were individually genotyped using microarrays. **b**, Genotype calls at ~52,000 markers for all four spores resulting from a single meiosis. Diagonal/vertical black lines are inferred crossovers; horizontal lines are non-crossovers. **c**–**e**, Close-ups of a crossover overlapped by an independent non-crossover in a third spore (**c**); a crossover with a complex gene conversion tract, and a nearby, independent non-crossover (**d**); and a long non-crossover at the end of the chromosome (**e**). (See Methods for full annotation procedure.) In close-ups, orange vertical segments denote markers with non-mendelian ratios (1:3 or 0:4).

found at the end of chromosomes, suggesting that they could be the result of meiotic break-induced replication, as has been proposed for long non-crossover tracts at the *HIS4* locus[22] (Fig. 1e). Three also showed complete loss of allelic variation across all four meiotic products (4:0 segregation), consistent with either mitotic or complex meiotic events.

We also observed that 11.5% of the conversion tracts accompanying crossovers exhibited complex patterns of genotype change (Fig. 1d). A total of 11.1% had more than one genotype change on just one of the involved chromatids, and 0.4%, on both chromatids. Such tracts are predicted to result from the resolution of a double Holliday junction owing to multiple distinct patches of heteroduplex in a single crossover event[6], but they could also possibly result from mismatch repair alternating between conversion and restoration. 3.4% of single-chromatid non-crossover events were also detected to have complex conversion tracts.

Recombination hotspots

To estimate the local recombination rate along the genome, we counted the events overlapping each intermarker interval and adjusted for the size of the interval (see Methods, Fig. 2b and Supplementary Figs 8 and 9). This novel approach was necessary because recombination events typically overlapped multiple markers, making existing rate estimation methods designed for low-resolution data inappropriate. Recombination hotspots were defined as runs of contiguous intermarker intervals involved in more recombination events than expected under a homogeneous genomic rate ($P < 0.001$, see Methods and Supplementary Information). We identified hotspots for crossover, non-crossover and overall recombination activity separately. At the hottest of the 179 resulting overall recombination hotspots, 27.7% of spores showed observable evidence of involvement in a crossover or non-crossover event (58.7% of meioses). At the hottest crossover and non-crossover hotspots, 21.7% and 8.7% of spores showed observable evidence of a crossover or a non-crossover, respectively. This corresponds to 21.7% and 17.4% of spores being involved in a crossover or non-crossover event, because a single crossover produces two spores with observable evidence whereas a non-crossover produces only one. Given that some non-crossovers may have been missed, we therefore observed similar rates for both outcomes at their hottest locations in the genome.

It is known that most DSBs occur in promoter regions[10,11], and indeed, 84% of hotspots overlap a promoter. Nonetheless, hotspot intervals primarily overlap coding sequence: only 25% of the bases in hotspot intervals overlap promoters, whereas 68% overlap coding sequences.

Centromere-proximal regions showed low recombination rates, and no recombination event overlapped a centromere on any chromosome (Supplementary Figs 8 and 9 and Supplementary Table 1). However, many chromosomes did have at least one event less than 4 kb away, including a crossover only 341 bp from *CEN5* (Supplementary Table 1). Telomeres could not be directly interrogated owing to repetitive sequence. We did, however, observe some chromosomes with a complete lack of recombination activity well before the telomeres; others showed strong activity near a telomere (Supplementary Figs 8 and 9).

Validating our approach, all previously known yeast recombination hotspots except for *HIS2* are within or adjacent to one of our hotspots (*HIS4*, *ARG4*, *CYS3*, *DED81*, *ARE1-IMG1*, *CDC19*, *THR4*, *LEU2-CEN3*)[23]. Furthermore, despite differences in strain background and the numerous heterozygosities in our hybrid strain, our recombination rates are in close agreement with a recently generated genome-wide DSB rate map in a homozygous SK1 strain[13] (Fig. 3). In addition to showing correspondence between the initiation of recombination and its resolution, this agreement suggests that the distribution of meiotic recombination is largely persistent within a species. Some fine-scale differences, however, do exist, possibly reflecting within-species variation in recombination rate[18].

Distinct crossover and non-crossover distributions

It is expected that the distribution of meiotic recombination is determined by the location of initiating DSBs as well as by how the DSBs are repaired[4]. It has not been clear, however, whether crossovers and

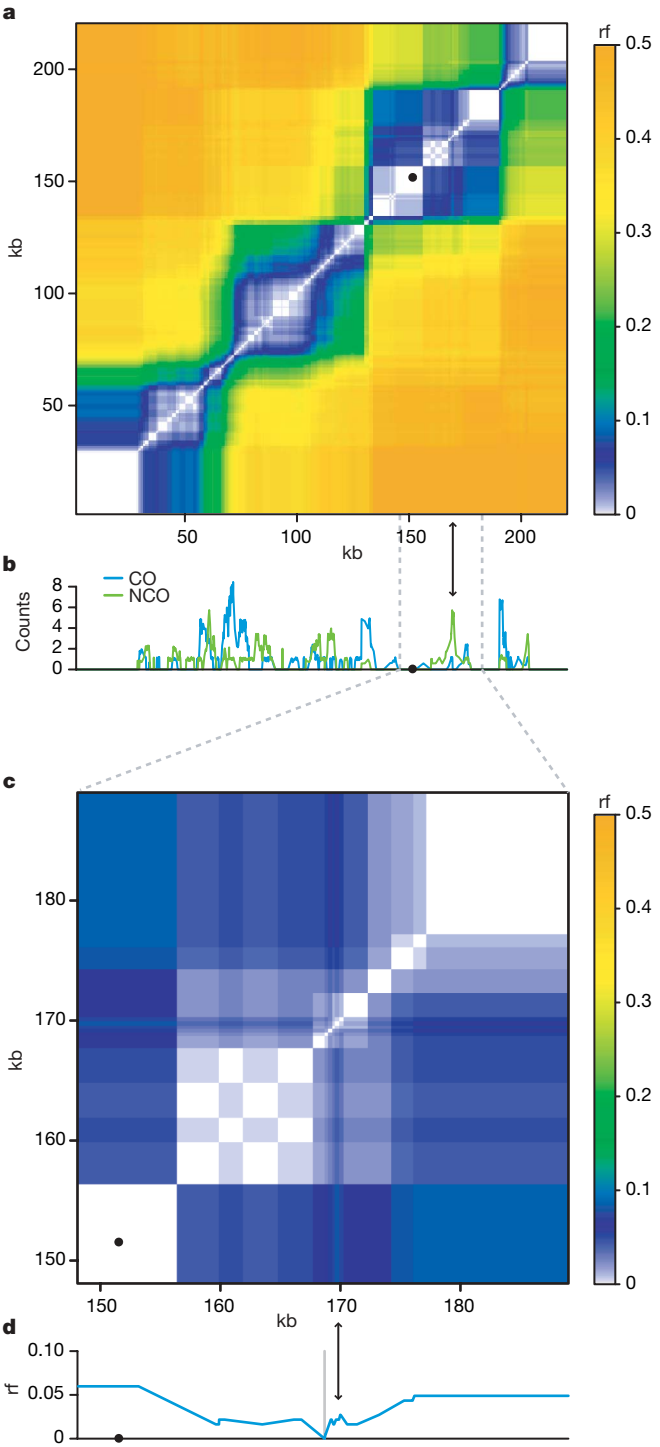


Figure 2 | Crossover and non-crossover rates along chromosome I and their effect on recombination fraction (rf). **a**, The recombination fraction was calculated for every pair of markers as the portion of segregants in which the markers have opposite genotype. Black dots denote the centromere. **b**, Crossover (CO, blue) and non-crossover (NCO, green) counts, adjusted for varying intermarker interval size (see Methods). **c**, Close-up of a non-crossover-biased region (indicated by arrows) on chromosome I. **d**, Recombination fraction relative to a single reference marker (grey vertical line) upstream of the non-crossover-biased region, showing the non-monotonic relationship between genetic and physical maps caused by such a region.

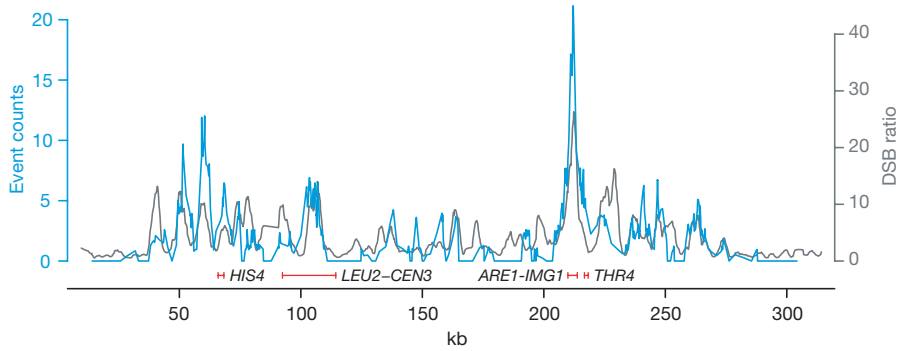


Figure 3 | Comparison of DSB and recombination rates along chromosome III. DSB smoothed fluorescence ratios in a SK1 strain (*dmc1*Δ, grey)[13] are compared with our recombination event counts (blue), after adjusting the latter for varying intermarker interval size (see Methods). Peak locations largely agree despite distinct strain backgrounds, although some fine-scale differences exist. Previously known hotspots are indicated by red segments.

non-crossovers always occur in similar proportions or whether there are crossover- or non-crossover-specific hotspots. Whereas a recent study reported mild crossover/non-crossover differences for two hotspots[24], our maps allow investigation of such differences genome wide (Fig. 2b). Using an approach that accounts for unobserved non-crossovers which fall completely between two markers, we identified regions with biased crossover/non-crossover ratios, and found more intermarker intervals with extreme ratios than expected by chance ($P < 0.0005$, see Methods). We observed an average excess of ~60 intervals favouring crossovers and ~170 intervals favouring non-crossovers, spanning ~100 kb of genomic sequence in total (see Methods). Notably, we estimated that such differences affect at least 1.4% of the genomic regions exhibiting one or more recombination event. The crossover/non-crossover event ratios at the regions showing the strongest evidence of bias, after accounting for the effect of marker spacing, were 14:0 and 0:7. Our findings therefore suggest that a significant fraction of the genome exhibits differences in crossover/non-crossover ratio.

The observed dissimilarity in crossover/non-crossover distribution has implications for linkage analysis. In contrast to crossover hotspots, regions with high non-crossover frequency can be expected to have reduced linkage to their surroundings, but to maintain linkage between loci to either side. By estimating the recombination fraction between all pairs of markers on each chromosome, we show that crossover hotspots are associated with linkage block boundaries, whereas non-crossover-biased regions correspond to regions with reduced linkage within blocks (Fig. 2). Non-crossover-biased regions result in a non-monotonic relationship between the genetic and physical distance, and create holes within linkage blocks. Over generations, non-crossover-biased hotspots would form genomic regions with low linkage disequilibrium relative to their surroundings, and thus be difficult to track with markers outside the non-crossover-biased region[25,26].

The existence of regions with a crossover/non-crossover bias suggests that the bifurcation between the two outcomes might, in fact, be a controlled process, influenced by local chromosomal properties. Recombination hotspots were found to contain short poly(A) stretches (20–41 bp) more frequently than expected, and to be significantly associated with several gene ontology (GO) terms (see Supplementary Information). Nonetheless, we found no sequence motifs to be specifically associated with crossover- or non-crossover-biased regions, and only one GO term ('cell ageing') to exhibit a significant association with such regions. A comparison of our results with measurement of transcriptional activity during meiosis in W303 and SK1 strains[27] showed that hotspot-proximal genes were significantly enriched in two specific expression profiles: a transcription peak around 2 h after meiotic induction ($P < 0.0001$, see Supplementary Information, Fig. 4a), and a transcription decrease between 8 and 10 h ($P = 0.0046$, Fig. 4b). In addition, a cluster with genes upregulated 4 h after meiotic induction contained genes from non-crossover-biased regions, but no genes from crossover-biased regions (Fisher exact test $P = 0.015$, Fig. 4c). This relationship between specific transcriptional behaviour and proximity to recombination hotspots supports a role for chromatin accessibility and transcription factor binding in meiotic recombination[28].

Crossovers and non-crossovers in recombination mutants

To assess the differences between the generation of crossovers and non-crossovers further, we mapped recombination events in *msh4* and *mms4* null mutants, in which either the Msh4/Msh5-dependent or the Mus81/Mms4-dependent crossover pathway is disturbed[7]. Five full tetrads of the *msh4* mutant were genotyped. Given the role of *MSH4* in crossover generation, its deletion is expected to reduce the number of crossovers but maintain the number of non-crossovers[29]. Consistent with this expectation, we observed that the non-crossover frequency showed no statistically significant change (*t*-test, $P = 0.12$), whereas the average number of crossovers per meiosis was markedly reduced from 90.5 in the wild type to 46.6 in *msh4* (*t*-test, $P < 0.0001$, Fig. 5a). Furthermore, in contrast to the wild type, all *msh4* tetrads except one had one or more chromosomes with no crossovers at all (6.3% of all chromosomes). Unexpectedly, the median size of *msh4* crossover conversion tracts was 479 bp larger than for wild type (Wilcoxon rank-sum, $P = 0.0003$). The median size of *msh4* non-crossover recombination tracts, however, was 338 bp shorter than for wild type (Wilcoxon rank-sum, $P = 0.0008$). Therefore, deletion of *MSH4* reduced genome-wide frequency of crossovers, as expected given its role in the Msh4/Msh5-dependent pathway, but affected tract size of both crossovers and non-crossovers (Supplementary Fig. 6).

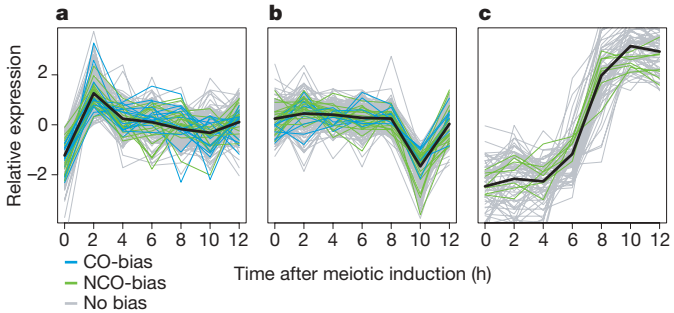


Figure 4 | Association between gene expression and recombination activity. After centring by their mean over time points, genes were clustered (see Supplementary Information) by their W303 expression profile during meiosis[27]. **a**, **b**, Two clusters enriched for genes overlapped by overall recombination hotspots. Genes overlapped by regions of crossover (CO) or non-crossover (NCO) bias, however, are present in similar proportion in these clusters. **c**, Gene cluster containing genes overlapped by regions with non-crossover bias, but no genes overlapped by regions with crossover bias.

The observation that, in the *msh4* mutant, the frequency of one event type was altered with respect to wild type whereas the other was not has two important implications. First, because Msh4 is thought to function downstream of DSB formation[30], we expect the *msh4* null mutant to have the same number of DSBs as the wild type. (This is known to be the case for *MSH5*, the functional partner of *MSH4* (ref. 31).) Our data therefore suggest that a fraction of DSBs are not resolved towards crossovers or non-crossovers, but may instead be repaired by alternative mechanisms such as sister chromatid exchange[32] or non-homologous end joining[4]. Second, we have perturbed a DSB-resolution pathway and seen strong but distinct effects on the global crossover/non-crossover balance. If this pathway has regional preferences, this may contribute to observed crossover/non-crossover bias.

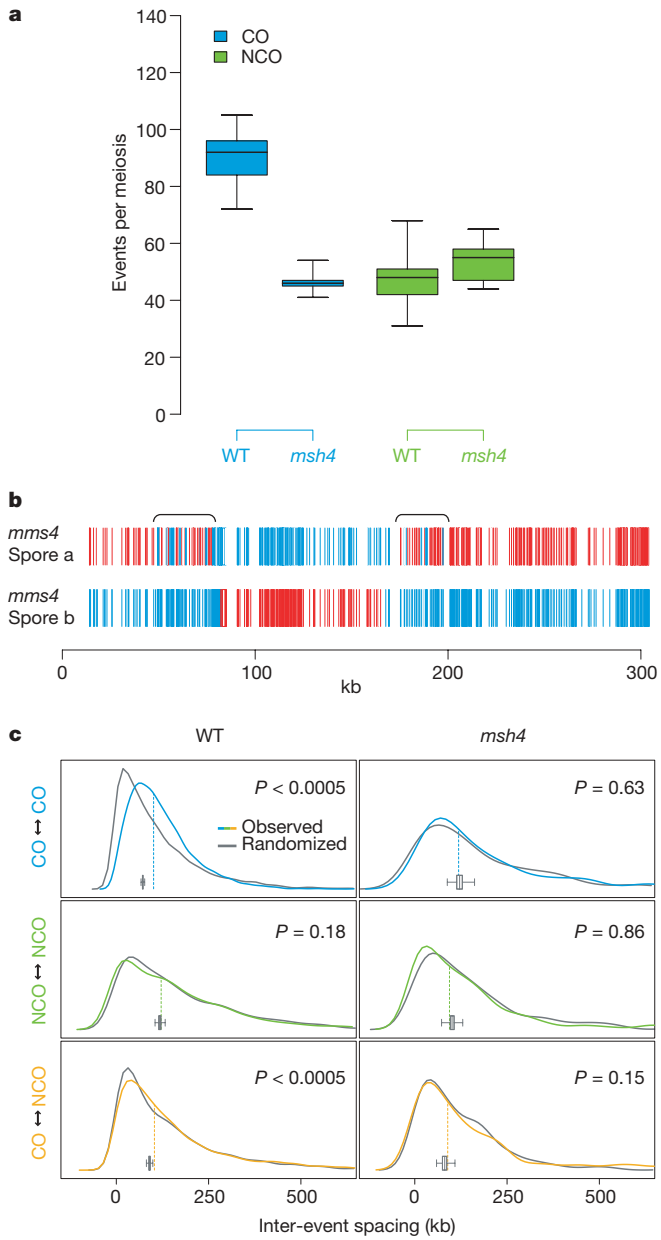


Figure 5 | Meiotic recombination in *msh4* and *mms4* strains. a, Genomic frequencies of crossovers (CO) and non-crossovers (NCO), per meiosis (wild type $n = 46$, *msh4* $n = 5$; box-plots show minimum, first quantile, median, third quantile and maximum). **b**, Genotype calls (S288c/YJM789, red/blue) along chromosome III in an *mms4* dyad showing large regions of frequent genotype change (calls in brackets validated by sequencing). **c**, Density estimates for the distance between adjacent recombination events, to measure interference. Coloured lines show the real data distribution, whereas grey lines denote one corresponding tetrad-randomized distribution. Dashed vertical lines show the real data median, and the box-plot shows the distribution of medians for 2,000 randomizations. *P*-values for difference in median distance are given within each panel (see Supplementary Information).

The *mms4* mutant exhibited low sporulation efficiency and spore viability, which impeded recovery of complete tetrads, so we only genotyped 6 dyads (12 spores) and 8 single *mms4* spores. Surprisingly, the *mms4* spores showed several regions (~7 per spore) exhibiting unusually frequent genotype changes (Fig. 5b)—up to ~70 kb in size and typically associated with apparent crossovers. For example, one such 63-kb region contained a total of 31 genotype changes. The mechanism responsible for these genotype changes is not known, but their presence may help elucidate the way in which the Mus81–Mms4 nuclease complex generates crossovers[8]. We chose not to pursue recombination event inference for the *mms4* spores owing to both the presence of such regions and the inherent difficulty in distinguishing between single non-crossovers and pairs of nearby crossovers in a single-spore context.

Crossover and non-crossover interference

Interference, where a recombination event reduces the probability that an additional recombination event occurs nearby[33], is an important determinant of the distribution of meiotic recombination, and could also contribute to differences in crossover/non-crossover rates. So far, interference has been reported only between crossovers[34]. To assess interference, we considered the distances between adjacent, same-tetrad recombination events. These distances were compared with those in tetrad-randomized data sets (see Supplementary Information). Tetrad randomization preserves hotspot and cold-spot structure along the genome, but removes interference effects. The distance between consecutive crossovers was larger in wild-type meioses than expected by chance: a median inter-crossover distance of 101.1 kb in observed data versus 71.8 kb under tetrad randomization ($P < 0.0005$, see Supplementary Information and Fig. 5c). No such effect was seen for non-crossovers. Notably, and in contrast to previous reports[34], crossovers and non-crossovers also exhibited interference: the median observed distance from a crossover to the nearest non-crossover was 13.1 kb larger in real data than under tetrad randomization ($P < 0.0005$). In the *msh4* null mutant, crossovers did not show interference ($P = 0.63$). This is consistent with the hypothesis that only crossovers generated by the Msh4/Msh5-dependent pathway exhibit interference[7]. Furthermore, in the *msh4* mutant, evidence of interference between crossovers and non-crossovers disappears as well ($P = 0.15$). These results support the existence of at least two types of crossovers with differences in interference, and yield genome-wide evidence for interference between crossovers, and among crossovers and non-crossovers.

We also observed an over-representation of overlapping events within the same meiosis in the wild-type strain, which is surprising given the observed patterns of interference. For example, 2.6% of crossover conversion tracts had an overlapping non-crossover partner on a third spore, and an additional 0.6% had an overlapping crossover partner involving the other two spores (Figs 1c and Supplementary Fig. 12). Such overlapping events could result from paired DSBs in two different chromatids; but, they could also be the consequence of a single DSB, the resolution of which involves multiple rounds of strand invasion and extension from different templates[35]. We also observed 110 pairs of partially or exactly overlapping non-crossovers with reciprocal genotypes. The existence of such pairs is relevant to current models for non-crossover formation (see Section 8 of Supplementary Information for discussion).

Genomic effect of gene conversion

Having observed differences in crossover and non-crossover distributions as well as interference between events, we next considered the effects of gene conversion tracts. We determined the portion of the

yeast genome that is involved in crossover-associated and non-crossover gene conversion. A total of 2.1% of the polymorphic positions was converted to the opposite genotype per meiosis. Furthermore, across the genomes of all four wild-type meiotic products, crossover tracts covered between 92 kb and 320 kb per meiosis (minimal and maximal), and the non-crossover tracts, between 62 kb and 148 kb. Therefore, as much as 1% of a meiotic product's genome may be subject to conversion in a single meiosis.

Genomic regions active in gene conversion are susceptible to the effect of gene conversion on allelic frequency, and also to mutation-prone processes[36]. We therefore analysed GC content and single-nucleotide polymorphism (SNP) density in converted regions and hotspots. For both crossover-associated and non-crossover gene conversions, we detected mismatch repair bias favouring GC nucleotides (Supplementary Information). Relative to the base content at SNP positions in the parental genomes, we observed a 1.4% GC increase in the converted sequences of the spores (χ^2 $P = 0.0001$, Supplementary Table 2). This bias could contribute to the association between recombination hotspots and GC-richness that we observed (χ^2 $P < 0.0001$)—an association that has also been found for DSBs[11]. Although on an evolutionary timescale, GC bias could potentially homogenize alleles, comparison to low-depth genome sequences of 37 *S. cerevisiae* strains showed that our hotspots were actually associated with greater genetic diversity (see Supplementary Information). Therefore, GC conversion bias may be counteracted by other processes, such as those that increase AT content[37,38]. We find no evidence of allelic homogenization at recombination hotspots, despite the presence of GC bias during mismatch repair.

Conclusion

The recombination maps presented here constitute the first survey of non-crossovers and both crossover-associated and non-crossover gene conversion across an entire genome in any organism. In addition to permitting detection and characterization of gene conversion, the high resolution of our approach reveals phenomena which would otherwise be difficult to observe, such as complex conversion tracts and large regions of frequent genotype changes (Figs 1d and 5b). The data uncover regions of interest for further investigation, and the approach is applicable to other mutants and conditions. It could thus contribute to answering questions about the mechanisms of interference and crossover homeostasis[24,39], or possible alternative DSB-resolution pathways[4–6].

Although the degree of polymorphism between the parental strains results in unprecedented marker resolution, polymorphisms may also affect recombination propensity[40,41]. Nonetheless, several observations suggest that recombination is not markedly perturbed in our hybrid: the agreement between our maps and the DSB map from a homozygous SK1 strain[13]; consistency between our overall number of crossovers and the number generated from genetic-map estimates[42]; and the detection of previously known recombination hotspots[23]. Furthermore, outside laboratory conditions, most sexually reproducing organisms are heterozygous. Individuals in natural populations may, therefore, resemble our hybrid more than they do a homozygous strain.

Our maps show the existence of locations with distinct preferences for either crossovers or non-crossovers, suggesting a role for genomic position in determining DSB resolution outcome. Given that chromatin conformation is known to be important for recombination generally[28], it is plausible that local chromosomal properties could influence the crossover/non-crossover bifurcation. Such properties may not, however, be the sole determinants of crossover/non-crossover bias. Through interference, both crossover–crossover and crossover–non-crossover, the decision could also depend on recombination activity in nearby regions.

Our maps also stress the relevance of non-crossovers, and gene conversion generally, in genetic analysis. Crossover is the major determinant of linkage disequilibrium, but both crossover-associated and non-crossover gene conversion weaken linkage disequilibrium between nearby loci. Models that incorporate gene conversion will therefore be able to relate linkage disequilibrium and physical distance more accurately. Furthermore, crossover-associated and non-crossover conversion tracts have different effects on the fine structure of haplotypes[26]. As shown in Fig. 2, gene conversion at crossover hotspots softens the boundaries of linkage blocks, whereas non-crossover-biased regions create holes within blocks. Both phenomena have implications for genetic association analyses. Although these regions are highly localized and have an impact on only a fraction of meioses, their effect can accumulate over generations, hiding genetic variants with phenotypic relevance (for example, disease genes). Having a higher density of markers in regions with frequent gene conversion may thus help to uncover genetic factors contributing to phenotypic variation.

METHODS SUMMARY

A S96/YJM789 hybrid strain was sporulated[43], and genomic DNA—from 51 wild type and 5 *msh4* tetrads as well as from 20 *mms4*, 13 S96 parental, and 12 YJM789 parental spores—was extracted from single-colony cultures and hybridized to a custom-designed tiling microarray[44]. (S96 is isogenic to S288c (refs 16, 17).) Normalized[45] fluorescence intensities corresponding to the set of probes covering each polymorphism were analysed by applying multivariate semi-supervised clustering to the combined parental and segregant data. Segregant genotypes were assigned using posterior probability of class membership. To reduce genotyping errors, we applied filters to whole arrays, to probe sets and to individual genotype calls. DNA sequencing of ~60 kb confirmed 100% of filtered genotype calls. After grouping data by tetrad, pairs of adjacent genotype change points isolated from all other changes were called non-crossovers if they involved one spore, or crossovers if they involved two. Complex groups of genotype changes were annotated as described in Supplementary Fig. 3. To calculate event rate along the genome, it was necessary to adjust for varying intermarker interval size. Because individual recombination events typically overlapped multiple intermarker intervals, a novel adjustment procedure was used (Supplementary Information). We defined three types of hotspots—crossover, non-crossover and overall recombination events—by identifying runs of contiguous intermarker intervals involved in more recombination events than expected under a homogeneous genomic rate. To assess crossover/non-crossover bias, we compared the number and size of intermarker intervals exhibiting more/fewer crossovers than expected to the corresponding null distribution, generated via simulation. We tested for interference—between consecutive events of the same type and also between crossovers and non-crossovers—by comparing the median distance between adjacent, same-tetrad events to medians computed after tetrad label randomization. This randomization strategy preserved hot- and cold-spot structure but removed interference.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 10 March; accepted 30 May 2008.
Published online 9 July 2008.

1. Gordo, I. & Charlesworth, B. Genetic linkage and molecular evolution. *Curr. Biol.* **11,** R684–R686 (2001).
2. Chen, J. M. *et al.* Gene conversion: mechanisms, evolution and human disease. *Nature Rev. Genet.* **8,** 762–775 (2007).
3. Page, S. L. & Hawley, R. S. Chromosome choreography: the meiotic ballet. *Science* **301,** 785–789 (2003).
4. Baudat, F. & de Massy, B. Regulating double-stranded DNA break repair towards crossover or non-crossover during mammalian meiosis. *Chromosome Res.* **15,** 565–577 (2007).
5. Bishop, D. K. & Zickler, D. Early decision; meiotic crossover interference prior to stable strand exchange and synapsis. *Cell* **117,** 9–15 (2004).
6. Whitby, M. C. Making crossovers during meiosis. *Biochem. Soc. Trans.* **33,** 1451–1455 (2005).
7. Argueso, J. L., Wanat, J., Gemici, Z. & Alani, E. Competing crossover pathways act during meiosis in *Saccharomyces cerevisiae*. *Genetics* **168,** 1805–1816 (2004).
8. Hollingsworth, N. M. & Brill, S. J. The Mus81 solution to resolution: generating meiotic crossovers without Holliday junctions. *Genes Dev.* **18,** 117–125 (2004).
9. Allers, T. & Lichten, M. Differential timing and control of noncrossover and crossover recombination during meiosis. *Cell* **106,** 47–57 (2001).
10. Baudat, F. & Nicolas, A. Clustering of meiotic double-strand breaks on yeast chromosome III. *Proc. Natl Acad. Sci. USA* **94,** 5213–5218 (1997).
11. Gerton, J. L. *et al.* Inaugural article: global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **97,** 11383–11390 (2000).

12. Borde, V. *et al.* Association of Mre11p with double-strand break sites during yeast meiosis. *Mol. Cell* **13,** 389–401 (2004).
13. Buhler, C., Borde, V. & Lichten, M. Mapping meiotic single-strand DNA reveals a new landscape of DNA double-strand breaks in *Saccharomyces cerevisiae. PLoS Biol.* **5,** 2797–2808 (2007).
14. Blitzblau, H. G. *et al.* Mapping of meiotic single-stranded DNA reveals double-stranded-break hotspots near centromeres and telomeres. *Curr. Biol.* **17,** 2003–2012 (2007).
15. Cherry, J. M. *et al.* Genetic and physical maps of *Saccharomyces cerevisiae. Nature* **387** (suppl.), 67–73 (1997).
16. McCusker, J. H., Clemons, K. V., Stevens, D. A. & Davis, R. W. Genetic characterization of pathogenic *Saccharomyces cerevisiae* isolates. *Genetics* **136,** 1261–1269 (1994).
17. Mortimer, R. K. & Johnston, J. R. Genealogy of principal strains of the yeast genetic stock center. *Genetics* **113,** 35–43 (1986).
18. Coop, G. *et al.* High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science* **319,** 1395–1398 (2008).
19. Borts, R. H. & Haber, J. E. Length and distribution of meiotic gene conversion tracts and crossovers in *Saccharomyces cerevisiae. Genetics* **123,** 69–80 (1989).
20. Jeffreys, A. J. & May, C. A. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nature Genet.* **36,** 151–156 (2004).
21. Terasawa, M. *et al.* Meiotic recombination-related DNA synthesis and its implications for cross-over and non-cross-over recombinant formation. *Proc. Natl Acad. Sci. USA* **104,** 5965–5970 (2007).
22. Merker, J. D., Dominska, M. & Petes, T. D. Patterns of heteroduplex formation associated with the initiation of meiotic recombination in the yeast *Saccharomyces cerevisiae. Genetics* **165,** 47–63 (2003).
23. Lichten, M. & Goldman, A. S. Meiotic recombination hotspots. *Annu. Rev. Genet.* **29,** 423–444 (1995).
24. Martini, E., Diaz, R. L., Hunter, N. & Keeney, S. Crossover homeostasis in yeast meiosis. *Cell* **126,** 285–295 (2006).
25. Ardlie, K. *et al.* Lower-than-expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion. *Am. J. Hum. Genet.* **69,** 582–589 (2001).
26. Wall, J. D. Close look at gene conversion hot spots. *Nature Genet.* **36,** 114–115 (2004).
27. Primig, M. *et al.* The core meiotic transcriptome in budding yeasts. *Nature Genet.* **26,** 415–423 (2000).
28. Petes, T. D. Meiotic recombination hot spots and cold spots. *Nature Rev. Genet.* **2,** 360–369 (2001).
29. Ross-Macdonald, P. & Roeder, G. S. Mutation of a meiosis-specific MutS homolog decreases crossing over but not mismatch correction. *Cell* **79,** 1069–1080 (1994).
30. Kunz, C. & Schar, P. Meiotic recombination: sealing the partnership at the junction. *Curr. Biol.* **14,** R962–R964 (2004).
31. Borner, G. V., Kleckner, N. & Hunter, N. Crossover/noncrossover differentiation, synaptonemal complex formation, and regulatory surveillance at the leptotene/zygotene transition of meiosis. *Cell* **117,** 29–45 (2004).
32. Schwacha, A. & Kleckner, N. Interhomolog bias during meiotic recombination: meiotic functions promote a highly differentiated interhomolog-only pathway. *Cell* **90,** 1123–1135 (1997).
33. Hillers, K. J. Crossover interference. *Curr. Biol.* **14,** R1036–R1037 (2004).
34. Malkova, A. *et al.* Gene conversion and crossing over along the 405-kb left arm of *Saccharomyces cerevisiae* chromosome VII. *Genetics* **168,** 49–63 (2004).
35. Oh, S. D. *et al.* BLM ortholog, Sgs1, prevents aberrant crossing-over by suppressing formation of multichromatid joint molecules. *Cell* **130,** 259–272 (2007).
36. Hurles, M. How homologous recombination generates a mutable genome. *Hum. Genomics* **2,** 179–186 (2005).
37. Birdsell, J. A. Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Mol. Biol. Evol.* **19,** 1181–1197 (2002).
38. Lindahl, T. Instability and decay of the primary structure of DNA. *Nature* **362,** 709–715 (1993).
39. Kleckner, N. *et al.* A mechanical basis for chromosome function. *Proc. Natl Acad. Sci. USA* **101,** 12592–12597 (2004).
40. Borts, R. H. & Haber, J. E. Meiotic recombination in yeast: alteration by multiple heterozygosities. *Science* **237,** 1459–1465 (1987).
41. Chen, W. & Jinks-Robertson, S. The role of the mismatch repair machinery in regulating mitotic and meiotic recombination between diverged sequences in yeast. *Genetics* **151,** 1299–1313 (1999).
42. Weiner, B. M. & Kleckner, N. Chromosome pairing via multiple interstitial interactions before and during meiosis in yeast. *Cell* **77,** 977–991 (1994).
43. Rockmill, B., Sym, M., Scherthan, H. & Roeder, G. S. Roles for two RecA homologs in promoting meiotic chromosome synapsis. *Genes Dev.* **9,** 2684–2695 (1995).
44. David, L. *et al.* A high-resolution map of transcription in the yeast genome. *Proc. Natl Acad. Sci. USA* **103,** 5320–5325 (2006).
45. Huber, W. *et al.* Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18** (suppl. 1), S96–S104 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank S. Clauder-Münster, M. Granovskaia, M. Sieber, T. Bähr-Ivacevic, M. Nguyen, V. Benes, Z. Xu, L. Ettwiller, P. McGettigan and the EMBL Genomics Core Facility for technical help; M. Knop for discussions; A. Akhtar, A. Ladurner, A. De Luna and M. Knop for critical comments on the manuscript; E. Louis, R. Durbin and D. Carter for making data from the Saccharomyces Genome Resequencing Project available; and the contributors to the Bioconductor (http://www.bioconductor.org) and R (http://www.R-project.org) projects for making their software available. This work was supported by grants to L.M.S. from the National Institutes of Health and the Deutsche Forschungsgemeinschaft, and to W.H. from the Human Frontier Science Program; and by a Darwin Trust's Jeff Shell Scholarship awarded to E.M.

Author Information Raw data are available from ArrayExpress (http://www.ebi.ac.uk/arrayexpress) under accession number E-TABM-470. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to L.M.S. (larsms@embl.de).

METHODS

Strains and media. The hybrid strain S288c/YJM789 was generated by crossing S96, isogenic to S288c (ref. 17), with YJM789[16]. To generate the homozygous *msh4*Δ and *mms4*Δ hybrid strains, the corresponding gene was replaced by a *nat*MX4 or *kan*MX4 drug-resistance marker[46] in each of the haploid parental strains, which were then crossed. Sporulation was induced by transferring overnight cultures from liquid YEPD to 2% potassium acetate[43].

DNA extraction and hybridization. Fifty-one complete wild-type and five complete *msh4* tetrads were dissected for genotyping. Twenty *mms4* viable spores were also selected, as were thirteen S96 and twelve YJM789 parentals. Spores were allowed to grow in YEPD solid medium and then streaked out to obtain single colonies, only one of which was used for genotyping. Note that starting from a single colony prevented analysis of heterozygosities within a single spore arising from post-meiotic segregation. Genomic DNA was extracted from an overnight, 100 ml, YEPD, saturated culture of each spore using a QIAGEN Genomic-tip according to manufacturer's protocol. Ten micrograms of genomic DNA were fragmented, biotin-labelled and hybridized to a custom Affymetrix microarray, as described previously[44]. All probes were remapped (exonerate[47]) to the S288c genome and the aligned portion of the YJM789 genome[48]. Only probes with one exact match (25 matching bases) and no near matches (22 to 24) were retained, yielding 287,000 S288c-specific probes, 112,000 YJM789-specific probes, and 2.37 million probes interrogating non-polymorphic sequence.

Genotyping. Fluorescence intensities were normalized with *vsn*[45,49]. SNPs, insertions and deletions were identified using the S288c/YJM789 alignment[48], and for each polymorphism, a probe set was formed from probes interrogating the position(s) involved. Nearby polymorphisms producing identical probe sets were treated as a single marker. Genotype labels were available for parental data, so to genotype segregants, semi-supervised clustering was applied to the combined parental and segregant data. For each probe set, a two-component gaussian mixture model—with fixed mixture proportions (0.5) but distinct covariance matrices—was fit using the EM algorithm. For the small fraction of probe sets with >10 probes (probe sets interrogating large indels), principal components dimension reduction ($d = 10$) was applied first. Segregant genotypes were assigned using posterior probability of class membership. For *mms4*, genotypes were assigned in a supervised fashion, using the distributions previously estimated from the wild-type and *msh4* data.

Filtering of genotype calls. We deliberately opted for a high no-call rate with fewer errors, to reduce the chance of spurious short non-crossovers. Five wild-type and two *mms4* arrays exhibiting excessive genotype switching and large Mahalanobis residuals were set aside. A small fraction (0.7%) of probe sets exhibiting >2 classes—probably due to cross-hybridization with unlinked loci—were discarded (Supplementary Fig. 2c). Misclassification rates were estimated using inferred mixture distributions, and probe sets (4.6%) for which this estimate exceeded 1% were also discarded (Supplementary Fig. 2b). For retained probe sets, individual calls (4.9%) were discarded if the posterior probability of assigned class membership was too far from 1, or if the Mahalanobis residual was large (Supplementary Fig. 2a). In two sequencing validation data sets covering ~60 kb—one focused on calls from 16 different wild-type spores, and another, on two regions of an *mms4* segregant exhibiting frequent genotype switching—100% of filtered genotype calls were confirmed.

Recombination event annotation. After collecting genotype data into tetrads, genotype change points were grouped by proximity. Most cases were simple: pairs of changes isolated from all other changes were called non-crossovers if they involved one spore, or crossovers if they involved two (Supplementary Fig. 3a). A fraction of cases, however, were more complex, admitting several distinct interpretations. To treat such cases systematically, cutoff-based rules reflecting basic assumptions about the recombination process were used. See Supplementary Information Section 1 for details. Importantly, we explored a variety of plausible alternative annotation sets, and found no qualitative change in our main results.

Conversion tract length. Tract size estimates obtained using midpoints of flanking intermarker intervals were used for most calculations (see Supplementary Information Section 3). Where indicated, we also computed lower and upper bounds, using the regions spanned by converted markers (minimal), and delimited by the two nearest unconverted markers (maximal)[2]. For summary statistics, we combined simple (Supplementary Fig. 3a) and complex (Supplementary Fig. 3b, c) conversion tracts.

Event rate adjustment. Intermarker interval size affects the probability of involvement in recombination events. To adjust for this, we used a semi-parametric statistical model (Supplementary Information) to relate size to the probabilities of (1) involvement in and (2) detection of recombination events. The model's extension length distribution was estimated empirically. Given this estimate, we then counted recombination events overlapping each intermarker interval, and estimated remaining parameters by Poisson regression.

Defining hotspots. Using model parameter estimates, expected crossover and non-crossover counts were computed for each intermarker interval under a null hypothesis of rate homogeneity. We identified three types of hotspots: crossover, non-crossover and overall recombination events. To identify crossover hotspots, we performed a one-tailed test ($\alpha = 0.001$) using the Poisson distribution and the expected crossover counts. Hot intermarker intervals separated by <500 bp were merged. Non-crossover and overall recombination hotspots were identified similarly. Note that the three types of hotspots are statistically related, but crossover and non-crossover hotspot counts need not sum to the overall count.

Crossover/non-crossover bias testing. To assess crossover/non-crossover bias, we used expected crossover and non-crossover counts to compute expected crossover fractions. Conditioning on the observed number of events overlapping each interval, we then compared observed and expected counts using two one-tailed binomial distribution tests. The resulting *P*-values correspond to either an excess or deficiency of crossovers. Despite the large sample size, individual intermarker intervals were rarely involved in >10 events, so we chose to treat crossover/non-crossover bias *P*-values collectively rather than individually. We simulated data ($B = 2000$) under the same binomial distributions used for *P*-value calculations—conditioning on observed counts so that rate inhomogeneity across the genome was preserved—and examined (1) the average number of simulated *P*-values falling below 0.10, and (2) the average total size of intermarker intervals associated with such *P*-values. The former permitted estimation of false discovery rate, and the latter, estimation of the total size of intermarker intervals associated with true crossover/non-crossover bias.

46. Goldstein, A. L. & McCusker, J. H. Three new dominant drug resistance cassettes for gene disruption in *Saccharomyces cerevisiae*. *Yeast* **15**, 1541–1553 (1999).
47. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
48. Wei, W. *et al.* Genome sequencing and comparative analysis of *Saccharomyces cerevisiae* strain YJM789. *Proc. Natl Acad. Sci. USA* **104**, 12825–12830 (2007).
49. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).

Vol 454|24 July 2008|doi:10.1038/nature07101

ARTICLES

Structure of a β_1-adrenergic G-protein-coupled receptor

Tony Warne[1], Maria J. Serrano-Vega[1], Jillian G. Baker[2], Rouslan Moukhametzianov[1], Patricia C. Edwards[1], Richard Henderson[1], Andrew G. W. Leslie[1], Christopher G. Tate[1] & Gebhard F. X. Schertler[1]

G-protein-coupled receptors have a major role in transmembrane signalling in most eukaryotes and many are important drug targets. Here we report the 2.7 Å resolution crystal structure of a β_1-adrenergic receptor in complex with the high-affinity antagonist cyanopindolol. The modified turkey (*Meleagris gallopavo*) receptor was selected to be in its antagonist conformation and its thermostability improved by earlier limited mutagenesis. The ligand-binding pocket comprises 15 side chains from amino acid residues in 4 transmembrane α-helices and extracellular loop 2. This loop defines the entrance of the ligand-binding pocket and is stabilized by two disulphide bonds and a sodium ion. Binding of cyanopindolol to the β_1-adrenergic receptor and binding of carazolol to the β_2-adrenergic receptor involve similar interactions. A short well-defined helix in cytoplasmic loop 2, not observed in either rhodopsin or the β_2-adrenergic receptor, directly interacts by means of a tyrosine with the highly conserved DRY motif at the end of helix 3 that is essential for receptor activation.

G-protein-coupled receptors (GPCRs) are a large family of integral membrane proteins that are prevalent in eukaryotes from yeast to man, and function as key intermediaries in the transduction of signals from outside to inside the cell[1]. Activating molecules (agonists), such as hormones and neurotransmitters, bind to GPCRs from the extracellular side of the cell membrane and induce a large conformational change that propagates to the cytoplasmic surface[2,3], resulting in activation of G proteins and a consequent change in the level of intracellular messengers such as cAMP, Ca^{2+} or signalling lipids. There are over 800 different human GPCRs[4], all of which share the characteristic arrangement of 7 transmembrane α-helices, with the polypeptide amino terminus on the extracellular side of the plasma membrane[5].

Analysis of the primary amino acid sequences of GPCRs has resulted in the definition of a number of families[6], the largest of which, family A, includes the archetypal GPCR, rhodopsin. The three human β-adrenergic receptor (βAR) subtypes, β_1, β_2 and β_3, belong to family A and share 51% sequence identity between $Trp^{1.31}$–$Asp^{5.73}$ and $Glu^{6.30}$–$Cys^{H8\text{-}Cterm}$; that is, excluding the amino and carboxy termini and most of cytoplasmic loop 3 (Supplementary Fig. 1; superscripts refer to Ballesteros–Weinstein numbering[7]). Drugs that inhibit β_1 and β_2 receptor signalling (antagonists and inverse agonists) are used to modulate heart function and are known as β-blockers[8], but selective β_1-antagonists are preferred because they have fewer side effects due to bronchial constriction by means of β_2 receptors in the lung. In contrast to the β_1 and β_2 receptors, the β_3-adrenergic receptor (β_3AR) is found in adipose tissue, where adrenaline stimulates metabolism, and is a potential target to treat obesity. Elucidation of the specificity determinants for drug affinity of the different βAR subtypes will allow the development of better subtype-specific β-blockers, with fewer side effects.

A milestone in the study of βARs was recently reached with the publication of a β_2-adrenergic receptor (β_2AR) structure in a complex with an antibody fragment, β_2AR–Fab[9], followed by the higher resolution structure of an engineered β_2AR fused in the middle of the third cytoplasmic loop (CL3) to T4 lysozyme, β_2AR–T4 (ref. 10). These structures, both containing the high affinity antagonist carazolol, defined the overall architecture of β_2AR and the structure of the ligand-binding pocket. However, the structures also raised questions of how a range of compounds can bind to the different but closely related βAR subtypes with different affinities. For example, the human β_1 and β_2 receptors are 67% identical within their transmembrane regions, but the residues that directly surround the ligand-binding pocket appear to be identical. Despite these similarities, larger antagonists such as CGP 20712A (see Supplementary Fig. 2) bind 500 times more strongly to β_1AR than to β_2AR, whereas ICI 118551 shows a 550-fold specificity for β_2AR over β_1AR[11]. There are also β_1- and β_2-specific agonists[12]. As an important step towards understanding subtype specificity, we have determined the structure of a β_1-adrenergic receptor (β_1AR).

Crystallization of β_1AR

GPCR crystallization is challenging, because GPCRs are usually unstable in detergent, contain unstructured regions and spontaneously cycle between an inactive antagonist state (*R*) and an active agonist state (*R**), which may further decrease the stability[13]. The human β_1AR is more difficult to purify than β_2AR because it is very unstable in detergent. We therefore used turkey (*M. gallopavo*) β_1AR, which is more stable than human β_1AR[14] although less stable than human β_2AR (M.J.S.-V. and C.G.T., unpublished observation). A mutated receptor, β_1AR-m23, was constructed with enhanced thermostability over the wild-type receptor and an altered equilibrium between *R* and *R** so that the mutant receptor was preferentially in the antagonist (*R*) state[15]. The receptor construct, β_1AR36-m23 (Fig. 1), purified in octylthioglucoside and in the presence of cyanopindolol gave good crystals showing isotropic diffraction beyond 2.7 Å.

Pharmacological analysis of β_1AR-m23

The mutant receptor β_1AR-m23 bound the antagonists dihydroalprenolol and cyanopindolol with similar affinities to the wild-type receptor, but the agonists noradrenaline and isoprenaline bound

[1]MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 0QH, UK. [2]Institute of Cell Signalling, Medical School, Queen's Medical Centre, University of Nottingham, Nottingham NG7 2UH, UK.

more weakly by a factor of 2,470 and 650, respectively[15]. This reflects a change in the R to R^* equilibrium of the receptor towards the antagonist R state. From this we predicted that, in a G-protein-coupling assay, the receptor would show no basal activity and that the concentration of agonist required for signalling would be orders of magnitude higher. Signalling assays were performed on stable cell lines expressing the wild-type β_1AR truncated at the N and C termini (β_1ARtrunc) and also containing the six thermostabilizing mutations (m23) (Supplementary Fig. 3). β_1ARtrunc-m23 coupled efficiently to G proteins and elicited a robust stimulation of cAMP-responsive reporter gene, although the agonist concentration response curve, as expected, was shifted to the right[16]. The drug ICI 118551, an inverse agonist for both β_1AR[17] and β_2AR[18], showed no reduction in the basal level of cAMP when added at a concentration 100-fold above its inhibition constant (K_i) to cells containing β_1ARtrunc-m23, implying there is negligible basal constitutive activity. The structure we have determined contains the very high affinity antagonist cyanopindolol in the binding pocket and represents closely the inactive conformation with respect to G-protein coupling.

Overall structure and the extracellular loops

The structure was solved by molecular replacement to 2.7 Å resolution with an R_{work} of 0.212 and an R_{free} of 0.268 (Supplementary Table 1). The four receptor molecules in the unit cell, labelled A–D (Supplementary Figs 4–6), were all very similar except that molecules A and D both had a 60° kink in helix 1 (H1). Also modelled were 31 water molecules, 4 Na^+ ions and 14 detergent molecules (see Supplementary Information). Unless otherwise stated, all further discussion refers to molecule B, because this molecule has an unkinked H1 and a relatively well-ordered H8. The helix boundaries, disordered regions and overall structural motifs are presented in Fig. 1.

The amino acid sequence of turkey β_1AR[19] is 82% and 67% identical to human β_1AR and human β_2AR, respectively, over residues Trp40$^{1.31}$–Asp242$^{5.73}$ and Glu285$^{6.30}$–Cys358$^{H8\text{-}Cterm}$ (that is, excluding the N and C termini and most of CL3); it is therefore expected that the structure of the transmembrane regions of β_1AR and β_2AR should be very similar. Our superposition of β_2AR (Protein Data Bank, PDB, code 2RH1) and β_1AR (chain B) is based on selected residues in H3, H5, H6 and H7 because we were particularly interested in comparing the ligand-binding pockets; 78 Cα atoms can be superimposed with a root mean square deviation (r.m.s.d.) of 0.25 Å. The r.m.s.d. over all transmembrane helices is 0.7 Å (269 Cα atoms; Supplementary Fig. 7). Comparison of the structures of β_1AR and β_2AR reveals no evidence for any significant changes in backbone conformation at the sites of the six point mutants introduced[15] to stabilize β_1AR. This is consistent with the observation that β_1AR-m23 binds antagonists with similar affinities to the wild-type receptor[15] and that it can couple efficiently to G proteins, although at higher agonist concentration (Supplementary Fig. 3). The basis for the thermostabilization by the six mutations R68$^{1.59}$S, M90$^{2.53}$V, Y227$^{5.58}$A, A282$^{6.27}$L, F327$^{7.37}$A and F338$^{7.48}$M is not immediately apparent from the structure.

The structures of the three extracellular loops (EL1–3) in β_1AR are very similar to those of β_2AR (Cα r.m.s.d. of 0.8 Å), consistent with the high sequence conservation of these regions in the βAR family (Supplementary Fig. 1). On the extracellular surface, a clear peak in the electron density is present at a position co-ordinated by the backbone carbonyl groups of residues Cys 192, Asp 195, Cys 198 and one or two water molecules (Supplementary Fig. 8). This density was assigned to a sodium ion on the basis of its coordination geometry[20]. Its role, bound at the negative end of the EL2 α-helix dipole, may be to stabilize the helical conformation of EL2 and thus the structure of the entrance to the ligand-binding pocket. The large difference in EL2 conformation between the α-helix found in β_2AR and the β-hairpin that closes off the retinal-binding site in rhodopsin is confirmed in the structure of β_1AR, suggesting that the α-helix may be a common feature in those GPCRs that bind their ligands rapidly and reversibly.

Cytoplasmic loop structure

In all GPCRs, CL2 and CL3 are believed to have an important role in the binding, selectivity and activation of G proteins, CL2 being

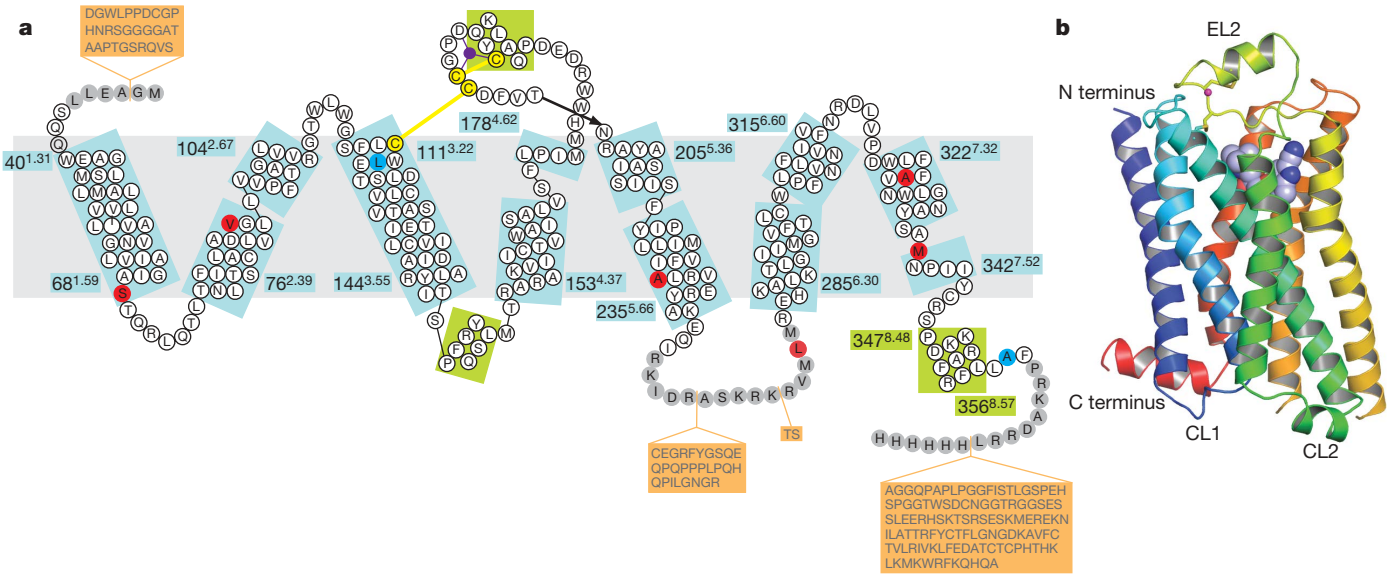


Figure 1 | Schematic representations of the turkey β_1AR structure. **a**, Diagram of the turkey β_1AR sequence in relation to secondary structure elements. The residues in white circles indicate regions that are well ordered; the sequences in grey circles were not resolved in the structure. The sequences on an orange background were deleted to make the β_1AR construct for expression. Thermostabilizing mutations are in red circles and two other mutations—C116L (increases functional expression) and C358A (eliminates palmitoylation site)—are in blue circles. The Na^+ ion is in purple. Numbers refer to the first and last amino acid residues in each helix (blue boxes), with the Ballesteros–Weinstein numbering in superscript. Helices were defined using the Kabsch and Sander algorithm[49], with helix distortions being defined as residues that have main chain torsion angles that differ by more than 40° from standard α-helix values (−60°,−40°). **b**, Ribbon representation of the β_1AR structure in rainbow colouration (N terminus, blue; C terminus, red), with the Na^+ ion in pink, the two near-by disulphide bonds in yellow, and cyanopindolol as a space-filling model. The extracellular loop 2 (EL2) and cytoplasmic loops 1 and 2 (CL1, CL2) are labelled.

important for the strength of the interaction and CL3 for specificity[21–25]. The β_1AR and β_2AR structures, along with rhodopsin[26], have similar conformations for CL1, but there are major differences in CL2 and CL3. The CL3 differences are not of physiological relevance because they arise from deletions (β_1AR), deletion and insertion of T4 lysozyme (β_2AR-T4) or formation of an antibody complex (β_2AR–Fab), with only the rhodopsin structure having a native CL3 (ref. 26). However, differences in the conformation of CL2 (Fig. 2) are important, because this region is very highly conserved between β_1AR and β_2AR, although poorly conserved with rhodopsin. In β_1AR, CL2 forms a short α-helix (residues Pro $146^{3.57}$–Leu $152^{3.63}$; Supplementary Fig. 9) parallel to the membrane surface whereas in both β_2AR structures and in rhodopsin this loop is in an extended conformation (Fig. 2). The α-helical conformation of CL2 observed in β_1AR cannot be accommodated in either the β_2AR–Fab complex[9] or the β_2AR–T4 fusion[10] crystal structures because of lattice contacts with adjacent molecules. In β_1AR, CL2 also makes lattice contacts, but these are different between each of the four molecules and it is therefore likely that the helical conformation found here represents the physiologically relevant structure for all βARs in the inactive conformation.

The CL2 loop has been proposed to function as the switch enabling G-protein activation[21], and it is clear from the β_1AR structure that this short α-helix interacts directly with the highly conserved Asp $138^{3.49}$Arg $139^{3.50}$Tyr $140^{3.51}$ (DRY) motif in H3. Tyr 149 in CL2 is located sufficiently close to Asp $138^{3.49}$ to allow the formation of a hydrogen bond (Fig. 2) between the tyrosine hydroxyl and the aspartate side chain. Supporting evidence for this structural role of Tyr 149 comes from the observation that the Y149A mutation makes β_1AR less thermally stable (Supplementary Table 2). The equivalent Tyr 141 in both β_2AR structures is in a cavity between H3, H4 and H6, but the biological relevance of this is unclear, owing to the perturbations in this region caused by either the T4 lysozyme fusion or by the bound antibody. Interestingly, a pattern of mutations consistent with an α-helical conformation for CL2 was found in the muscarinic M5 receptor, and the equivalent M5 mutation Y138A led to increased constitutive activity[21]. Thus, it is likely that both the tyrosine residue and the CL2 α-helix have key roles in G-protein coupling.

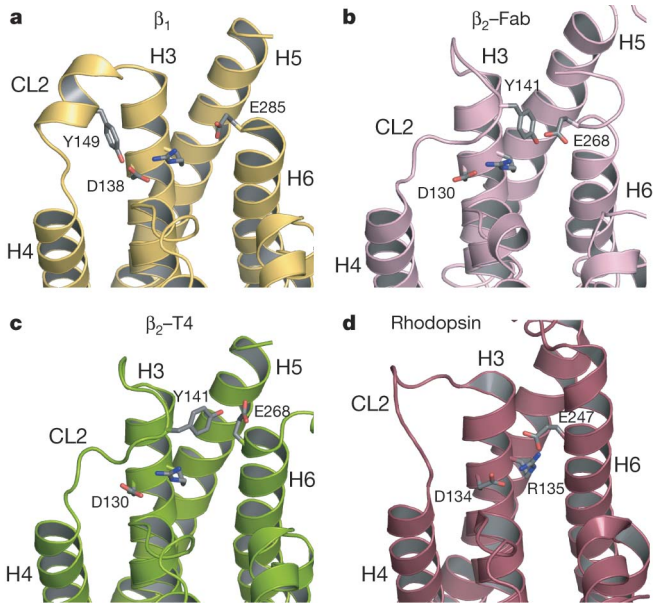


Figure 2 | Comparison of the CL2 loop regions in four GPCR structures. **a–d**, Shown are β_1AR (**a**), the β_2AR–Fab complex (**b**), the β_2AR–T4 lysozyme fusion (**c**) and rhodopsin (**d**). Residues DR from the highly conserved $D^{3.48}R^{3.49}Y^{3.50}$ motif are shown. Residue $E^{6.30}$, which is half of the putative ionic lock, is also shown as E247 in rhodopsin, and E285 and E268 in β_2AR, respectively: $E247^{6.30}$ was thought to form a salt bridge with $R135^{3.49}$ in rhodopsin, but the evidence is weak. Finally, Y149 in β_1 forms a hydrogen bond with $D138^{3.48}$ in β_1AR.

A salt bridge between $Arg^{3.50}$ and $Glu^{6.30}$, termed the 'ionic lock' (Fig. 2), was proposed to have an essential role in maintaining GPCRs in an inactive state[27] but to break on receptor activation. Because the β_1AR structure represents a receptor lacking basal activity and containing bound antagonist, it is highly likely to represent the *R* conformation. However, this salt bridge is not present in either the β_1AR or the β_2AR structures (Fig. 2). This suggests that the ionic lock is not an essential feature of the inactive state. Even in dark-state rhodopsin, where these two charged residues are within hydrogen bonding distance[26,28,29], the side chain B-factors of the two residues differ greatly (by 20–40 $Å^2$)[26] so there is no direct experimental evidence for any 'lock'.

Selectivity of the ligand-binding pocket

The two β-receptor antagonists, cyanopindolol and carazolol, have very similar chemical structures (Supplementary Fig. 2) and both ligands bind with very high affinity to all βARs. Carazolol is present in the ligand-binding pocket of both β_2 structures, whereas the structure of β_1 contains cyanopindolol. In the β_1AR structure there are 15 amino acid residues (using a 3.9 Å distance criterion) for which the side chains make contacts with cyanopindolol: 4 side chains are from H3, 3 are from H5, 4 are from H6, 2 are from H7 and 2 are from EL2 (Fig. 3). All of these residues are identical to those in human β_2AR, and the mode of binding of cyanopindolol to β_1AR is, therefore, similar to that of carazolol in β_2AR. However, the extra ring in the carazolol heterocyclic ring, owing to van der Waals contact with Tyr $199^{5.38}$ in β_2AR, pushes the ligand more deeply into the binding site. The nitrogen in the cyano-moiety of cyanopindolol makes a weak hydrogen bond with the hydroxyl of Thr $203^{5.34}$, which is located together with $Phe201^{5.32}$ on EL2 (Fig. 3). The same hydrogen bonds between the ligand and Asp $121^{3.32}$, Asn $329^{7.39}$ and Ser $211^{5.42}$ are present in both β_1AR and β_2AR structures, but the side-chain rotamer conformation of Ser $211^{5.42}$ is different (Fig. 4 and Methods).

To explain why some ligands preferentially bind to either β_1AR or β_2AR, which is important in understanding the sub-type specificity of the human receptors[11], there must be differences in amino acid residues close to the ligand-binding pocket that directly or indirectly affect binding. A comparison of residues within 8 Å of the binding pocket identified only two residues that are different between human β_1AR and β_2AR subtypes. The respective residues are Val $172^{4.56}$ and Phe $325^{7.35}$ in β_1AR, equivalent to Thr $164^{4.56}$ and Tyr $308^{7.35}$ in β_2AR. These differences introduce polar residues near the binding pocket of β_2AR relative to β_1AR (Fig. 4), which could affect ligand selectivity. Mutagenesis studies[30,31] have also implied that Tyr $308^{7.35}$ is important for agonist selectivity in β_2AR. In β_2AR, Tyr $308^{7.35}$ is positioned close to the binding pocket and can form a hydrogen bond to Asn $293^{6.55}$. In β_1AR the side chain of Asn $310^{6.55}$ is closer to the cyano group of cyanopindolol and the equivalent residue, Phe $325^{7.35}$, is further from the binding pocket (Fig. 4). As a result, there is no contact between Phe $325^{7.35}$ in β_1AR and cyanopindolol.

Part of the ligand-binding site is formed by EL2, and the backbone positions within this highly structured region of β_1AR differ from β_2AR by an r.m.s.d. of only 0.84 Å, compared with 0.63 Å between the same residues in molecules A and B in the unit cell. There are also significant differences in the primary amino acid sequence in this region that change the shape and charge distribution around the entrance to the ligand-binding pocket (Supplementary Fig. 10), with an ion pair formed between Asp $192^{5.31}$ and Lys $305^{7.32}$ in β_2AR that is absent in β_1AR because the respective residues are both aspartate (Asp $200^{5.31}$ and Asp $322^{7.32}$). Differences between β_1AR and β_2AR in this region could affect ligand binding, especially for larger ligands with extensions that have direct interactions with non-conserved side chains. Recent mutational studies show that EL2 influences the specificity of ligand binding to both the normal (orthosteric) site[32,33] and

the sites of allosteric modulators[34], and that the loop flexibility is important to the binding kinetics[35].

The structure of β_1AR, when compared to that of β_2AR, provides a sound basis for studying selectivity differences between βAR antagonists that are structurally similar to cyanopindolol and carazolol. However, many ligands, such as the inverse agonist CGP 20712A (Supplementary Fig. 2), show very high selectivities[11] but are physically larger and structurally distinct from either cyanopindolol or carazolol. These ligands could well make contact with residues other than those described here.

Agonist binding and GPCR activation

The β_1AR crystal structure shows the inactive state of the receptor, but it is notable that many agonists, including the natural ligands adrenaline and noradrenaline, are smaller than many of the best antagonists, including cyanopindolol. Agonists have a shorter distance, by two carbon–carbon bonds or 2–3 Å, between the catechol hydroxyl groups or their equivalent and the obligatory amine nitrogen. We superimposed (Fig. 4b) a model of adrenaline with that of cyanopindolol and examined its relationship to the side chains of Asp 121$^{3.32}$ and Asn 329$^{7.39}$, which make hydrogen bonds with the amine, and those of Ser 211$^{5.42}$, Ser 212$^{5.43}$ and Ser 215$^{5.46}$, which are expected to hydrogen bond with the meta- and para-hydroxyl groups on the catechol ring[36–38]. As noticed previously[39], the catechol hydroxyl groups are well spaced and well oriented to interact with the side chain hydroxyl groups of Ser 211$^{5.42}$, Ser 212$^{5.43}$ and Ser 215$^{5.46}$ on H5, but cannot reach far enough to make good hydrogen bonds if the amine occupies the same position as it does adjacent to Asp 121$^{3.32}$ in the cyanopindolol complex, without a substantial structural change in the receptor. It seems very reasonable that the

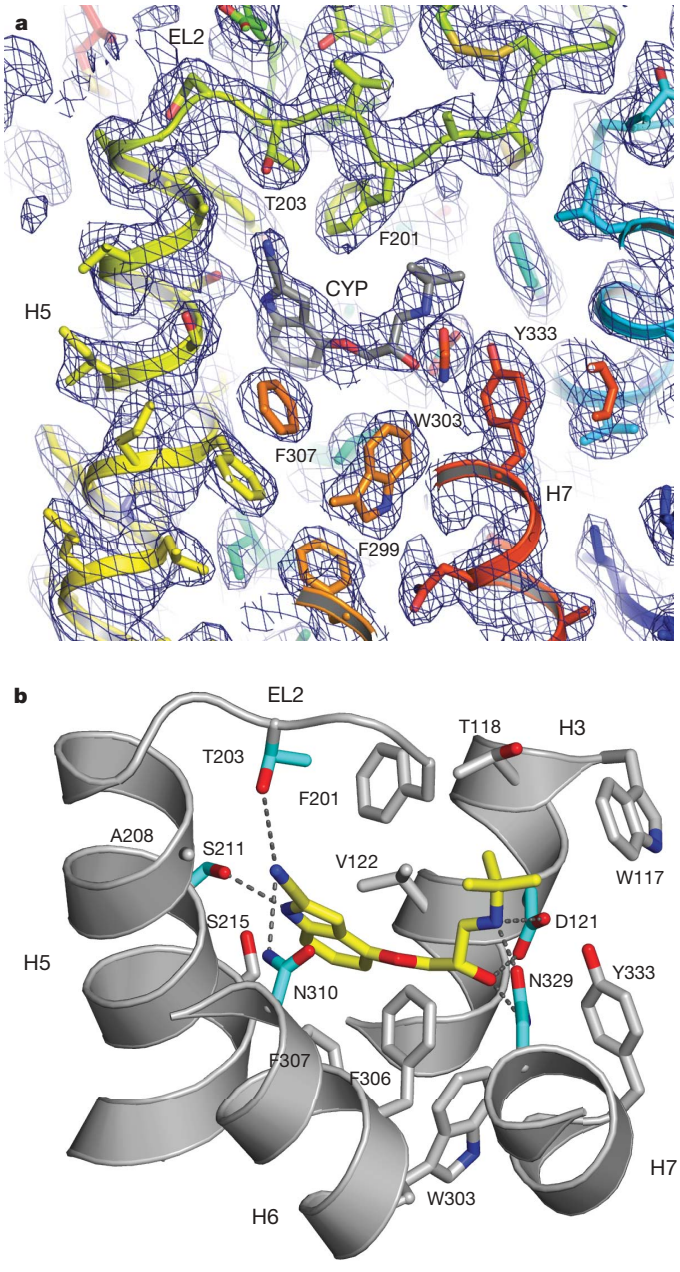


Figure 3 | Structure of the ligand-binding pocket. **a**, $2F_o$–F_c map before inclusion of cyanopindolol (CYP) in the model, showing the interaction of CYP with Thr 203 and Phe 201 in EL2. **b**, Amino acid residues that interact with the ligand cyanopindolol (yellow) by polar interactions (aquamarine) or non-polar interactions (grey).

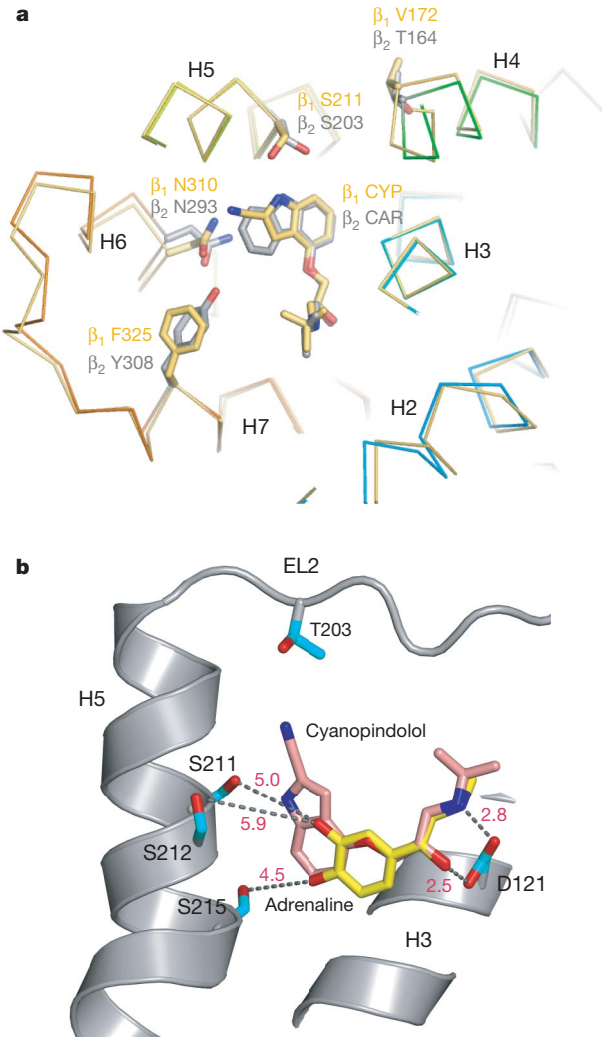


Figure 4 | Comparisons between β receptor ligand-binding pockets and the binding of different ligands. **a**, Superposition of β_1AR molecule B with β_2AR (PDB code 2RH1, ref. 10) in the region surrounding the ligand-binding site. Shown are side chains that have different rotamer conformations (N310$^{6.55}$ and S211$^{5.42}$) along with two residues that are conserved yet consistently different between β_1 and β_2 receptors (F325/Y308$^{7.35}$ and V172/T164$^{4.56}$). Cyanopindolol (CYP) is in the ligand-binding pocket of the β_1 receptor, and carazolol (CAR) is in the β_2 receptor. The biggest backbone deviation is seen at the V172/T164$^{4.56}$ position. **b**, Superposition of a model of the agonist, adrenaline (yellow), with the structure of the antagonist, cyanopindolol (pink), as it binds to β_1AR, showing the distances (in Å, red) to the nearest side chains known to interact with the hydroxyl groups on the catechol ring of the agonist. It is clear that a 2–3 Å tightening of the pocket around the ligand must occur on agonist binding.

ligand-binding site in β_1AR will contract by 2–3 Å on activation so that both ends of adrenaline can make good interactions with the residues on H3/H7 and H5. This view is also supported by engineered zinc-binding sites that activate the receptor[40,41]. How this tightening around the ligand-binding site could propagate to the cytoplasmic surface and cause an outward 5–6 Å movement of H6 (refs 2, 3) is difficult to predict, because all the transmembrane helices except H1 and H3 have pronounced kinks at conserved proline residues, which means they could easily bend. However, one speculation is that the pulling of H5 towards the centre of the receptor on activation could force H3 and H6 apart, causing cytoplasmic loops CL2 and CL3 to move apart, as observed in photoactivated rhodopsin[3], and trigger recruitment of the G-protein complex.

METHODS SUMMARY

Purification and crystallization. The β_1AR construct T34–424/His 6 (see ref. 42) was the starting point for the generation of the β_1AR36-m23 construct that crystallized. The C terminus was further truncated after Leu 367, and six histidines were added. Two segments, comprising residues 244–271 and 277–278 of CL3, were also deleted. The construct included the following eight point mutations: C116$^{3.27}$L increased expression; C358A at the C terminus of H8 removed palmitoylation and helped crystallization; and R68$^{1.59}$S, M90$^{2.53}$V, Y227$^{5.58}$A, A282$^{6.27}$L, F327$^{7.37}$A and F338$^{7.48}$M thermostabilized the receptor in the antagonist conformation[15]. The receptor was expressed using the baculovirus system and then purified[42] in decylmaltoside, with a detergent exchange to octylthioglucoside on the alprenolol sepharose column. Crystals were obtained by vapour diffusion at 18 °C with hanging drops after addition of an equal volume of reservoir solution (0.1 M *N*-(2-acetamido)iminodiacetic acid:NaOH, pH 6.9–7.3, and 29–32% PEG600) to purified receptor (6.0 mg ml^{-1}).

Data collection, structure solution and refinement. Diffraction data were collected from many crystals on beamlines ID13 and ID23-2 at ESRF, Grenoble[43,44]; the data used for structure determination were collected at ID23-2 with a 10 µm beam using three positions on a single cryo-cooled crystal (100 K). Images were processed with MOSFLM and SCALA[45]. The structure was solved by molecular replacement with PHASER[46], using the structure of human β_2AR[10] as an initial model. All four copies of the molecule in the triclinic unit cell were located (Supplementary Figs 4 and 5). The amino acid sequence was corrected, and the model refined with PHENIX[47] and rebuilt with O[48] (see Methods for further details). An overview of the B-factor distribution for β_1AR molecules A and B is shown in Supplementary Fig. 6. Figures were produced using Pymol (DeLano Scientific LLC).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 26 March; accepted 19 May 2008.
Published online 25 June 2008.

1. Fredriksson, R. & Schioth, H. B. The repertoire of G-protein-coupled receptors in fully sequenced genomes. *Mol. Pharmacol.* **67,** 1414–1425 (2005).
2. Hubbell, W. L., Altenbach, C., Hubbell, C. M. & Khorana, H. G. Rhodopsin structure, dynamics, and activation: a perspective from crystallography, site-directed spin labeling, sulfhydryl reactivity, and disulfide cross-linking. *Adv. Protein Chem.* **63,** 243–290 (2003).
3. Altenbach, C. *et al.* High resolution distance mapping in rhodopsin reveals the pattern of helix movement due to activation. *Proc. Natl Acad. Sci. USA* **105,** 7439–7444 (2008).
4. Foord, S. M. *et al.* International Union of Pharmacology. XLVI. G protein-coupled receptor list. *Pharmacol. Rev.* **57,** 279–288 (2005).
5. Baldwin, J. M., Schertler, G. F. & Unger, V. M. An alpha-carbon template for the transmembrane helices in the rhodopsin family of G-protein-coupled receptors. *J. Mol. Biol.* **272,** 144–164 (1997).
6. Bockaert, J. & Pin, J. P. Molecular tinkering of G protein-coupled receptors: an evolutionary success. *EMBO J.* **18,** 1723–1729 (1999).
7. Ballesteros, J. A. & Weinstein, H. Integrated methods for the construction of three dimensional models and computational probing of structure function relations in G protein-coupled receptors. *Methods Neurosci.* **25,** 366–428 (1995).
8. Black, J. W. Drugs from emasculated hormones — the principle of syntopic antagonism (Nobel lecture). *Angew. Chem. Int. Edn Engl.* **28,** 886–894 (1989).
9. Rasmussen, S. G. *et al.* Crystal structure of the human β_2 adrenergic G-protein-coupled receptor. *Nature* **450,** 383–387 (2007).
10. Cherezov, V. *et al.* High-resolution crystal structure of an engineered human β_2-adrenergic G protein-coupled receptor. *Science* **318,** 1258–1265 (2007).
11. Baker, J. G. The selectivity of β-adrenoceptor antagonists at the human β_1, β_2 and β_3 adrenoceptors. *Br. J. Pharmacol.* **144,** 317–322 (2005).
12. Sugimoto, Y. *et al.* β_1-selective agonist (-)-1-(3,4-dimethoxyphenetylamino)-3-(3,4-dihydroxy)-2-propanol [(-)-RO363] differentially interacts with key amino acids responsible for β_1-selective binding in resting and active states. *J. Pharmacol. Exp. Ther.* **301,** 51–58 (2002).
13. Gether, U. *et al.* Structural instability of a constitutively active G protein-coupled receptor. Agonist-independent activation due to conformational flexibility. *J. Biol. Chem.* **272,** 2587–2590 (1997).
14. Parker, E. M., Kameyama, K., Higashijima, T. & Ross, E. M. Reconstitutively active G protein-coupled receptors purified from baculovirus-infected insect cells. *J. Biol. Chem.* **266,** 519–527 (1991).
15. Serrano-Vega, M. J., Magnani, F., Shibata, Y. & Tate, C. G. Conformational thermostabilization of the beta1-adrenergic receptor in a detergent-resistant form. *Proc. Natl Acad. Sci. USA* **105,** 877–882 (2008).
16. Baker, J. G. Site of action of β-ligands at the human β_1-adrenoceptor. *J. Pharmacol. Exp. Ther.* **313,** 1163–1171 (2005).
17. Lattion, A., Abuin, L., Nenniger-Tosato, M. & Cotecchia, S. Constitutively active mutants of the β_1-adrenergic receptor. *FEBS Lett.* **457,** 302–306 (1999).
18. Samama, P. *et al.* Negative antagonists promote an inactive conformation of the β_2-adrenergic receptor. *Mol. Pharmacol.* **45,** 390–394 (1994).
19. Yarden, Y. *et al.* The avian β-adrenergic receptor: primary structure and membrane topology. *Proc. Natl Acad. Sci. USA* **83,** 6795–6799 (1986).
20. Harding, M. M. Metal-ligand geometry relevant to proteins and in proteins: sodium and potassium. *Acta Crystallogr. D* **58,** 872–874 (2002).
21. Burstein, E. S., Spalding, T. A. & Brann, M. R. The second intracellular loop of the m5 muscarinic receptor is the switch which enables G-protein coupling. *J. Biol. Chem.* **273,** 24322–24327 (1998).
22. Wong, S. K. F., Parker, E. M. & Ross, E. M. Chimeric muscarinic cholinergic β-adrenergic receptors that activate Gs in response to muscarinic agonists. *J. Biol. Chem.* **265,** 6219–6224 (1990).
23. Wong, S. K. F. & Ross, E. M. Chimeric muscarinic cholinergic:β-adrenergic receptors that are functionally promiscuous among G-proteins. *J. Biol. Chem.* **269,** 18968–18976 (1994).
24. Wess, J., Bonner, T. I., Dorje, F. & Brann, M. R. Delineation of muscarinic receptor domains conferring selectivity of coupling to guanine nucleotide-binding proteins and 2nd messengers. *Mol. Pharmacol.* **38,** 517–523 (1990).
25. Scarselli, M., Li, B., Kim, S. K. & Wess, J. Multiple residues in the second extracellular loop are critical for M3 muscarinic acetylcholine receptor activation. *J. Biol. Chem.* **282,** 7385–7396 (2007).
26. Li, J. *et al.* Structure of bovine rhodopsin in a trigonal crystal form. *J. Mol. Biol.* **343,** 1409–1438 (2004).
27. Ballesteros, J. A. *et al.* Activation of the β_2-adrenergic receptor involves disruption of an ionic lock between the cytoplasmic ends of transmembrane segments 3 and 6. *J. Biol. Chem.* **276,** 29171–29177 (2001).
28. Palczewski, K. *et al.* Crystal structure of rhodopsin: a G protein-coupled receptor. *Science* **289,** 739–745 (2000).
29. Okada, T. *et al.* The retinal conformation and its environment in rhodopsin in light of a new 2.2 angstrom crystal structure. *J. Mol. Biol.* **342,** 571–583 (2004).
30. Kikkawa, H., Isogaya, M., Nagao, T. & Kurose, H. The role of the seventh transmembrane region in high affinity binding of a β_2-selective agonist TA-2005. *Mol. Pharmacol.* **53,** 128–134 (1998).
31. Isogaya, M. *et al.* Identification of a key amino acid of the β_2-adrenergic receptor for high affinity binding of salmeterol. *Mol. Pharmacol.* **54,** 616–622 (1998).
32. Shi, L. & Javitch, J. A. The second extracellular loop of the dopamine D2 receptor lines the binding-site crevice. *Proc. Natl Acad. Sci. USA* **101,** 440–445 (2004).
33. Klco, J. M., Wiegand, C. B., Narzinski, K. & Baranski, T. J. Essential role for the second extracellular loop in C5a receptor activation. *Nature Struct. Mol. Biol.* **12,** 320–326 (2005).
34. Voigtlander, U. *et al.* Allosteric site on muscarinic acetylcholine receptors: Identification of two amino acids in the muscarinic M-2 receptor that account entirely for the M-2/M-5 subtype selectivities of some structurally diverse allosteric ligands in *N*-methylscopolamine-occupied receptors. *Mol. Pharmacol.* **64,** 21–31 (2003).
35. Avlani, V. A. *et al.* Critical role for the second extracellular loop in the binding of both orthosteric and allosteric G protein-coupled receptor ligands. *J. Biol. Chem.* **282,** 25677–25686 (2007).
36. Sato, T., Kobayashi, H., Nagao, T. & Kurose, H. Ser(203) as well as Ser(204) and Ser(207) in fifth transmembrane domain of the human β_2-adrenoceptor contributes to agonist binding and receptor activation. *Br. J. Pharmacol.* **128,** 272–274 (1999).
37. Strader, C. D. *et al.* Identification of 2 serine residues involved in agonist activation of the β-adrenergic-receptor. *J. Biol. Chem.* **264,** 13572–13578 (1989).
38. Liapakis, G. *et al.* The forgotten serine — A critical role for Ser-203(5.42) in ligand binding to and activation of the β_2-adrenergic receptor. *J. Biol. Chem.* **275,** 37779–37788 (2000).
39. Rosenbaum, D. M. *et al.* GPCR engineering yields high-resolution structural insights into β_2-adrenergic receptor function. *Science* **318,** 1266–1273 (2007).
40. Elling, C. E., Thirstrup, K., Holst, B. & Schwartz, T. W. Conversion of agonist site to metal-ion chelator site in the β_2-adrenergic receptor. *Proc. Natl Acad. Sci. USA* **96,** 12322–12327 (1999).
41. Schwartz, T. W. *et al.* Molecular mechanism of 7TM receptor activation — a global toggle switch model. *Annu. Rev. Pharmacol. Toxicol.* **46,** 481–519 (2006).

42. Warne, T., Chirnside, J. & Schertler, G. F. Expression and purification of truncated, non-glycosylated turkey beta-adrenergic receptors for crystallization. *Biochim. Biophys. Acta* **1610,** 133–140 (2003).
43. Riekel, C., Burghammer, M. & Schertler, G. Protein crystallography microdiffraction. *Curr. Opin. Struct. Biol.* **15,** 556–562 (2005).
44. Standfuss, J. *et al.* Crystal structure of a thermally stable rhodopsin mutant. *J. Mol. Biol.* **372,** 1179–1188 (2007).
45. Evans, P. Scaling and assessment of data quality. *Acta Crystallogr. D* **62,** 72–82 (2006).
46. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Cryst.* **40,** 658–674 (2007).
47. Adams, P. D. *et al.* PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. D* **58,** 1948–1954 (2002).
48. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron-density maps and the location of errors in these models. *Acta Crystallogr. A* **47,** 110–119 (1991).
49. Kabsch, W. & Sander, G. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22,** 2577–2637 (1983).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by a joint grant from Pfizer Global Research and Development and from the MRCT Development Gap Fund to C.G.T. and R.H., in addition to core funding from the MRC. G.F.X.S. was financially supported by a Human Frontier Science Project (HFSP) programme grant (RG/0052), a European Commission FP6 specific targeted research project (LSH-2003-1.1.0-1) and an ESRF long-term proposal. J.G.B. is funded by a Wellcome Trust Clinician Scientist Fellowship. We thank E. Ross for his support in the initial stages of the β_1AR project at the LMB and for his comments on the manuscript. In addition, we would also like to thank R. Grisshammer, E. Hulme, F. Marshall and M. Weir, as well as J. Li, M. Babu and other colleagues at LMB for their comments. We also thank beamline staff at the European Synchrotron Radiation Facility, particularly C. Riekel and M. Burghammer at ID13 and D. Flot and S. McSweeney at ID 23-2. Finally, we thank D. Loakes for Supplementary Fig. 2.

Author Contributions T.W. devised and carried out receptor expression, purification, crystallization and cryo-cooling of the crystals. Receptor stabilization and baculovirus expression were performed by M.J.S.-V.; both authors were also involved in data collection and preliminary crystallographic analyses of the crystals. P.C.E. helped with the crystal cryo-cooling strategy and in diffraction data collection. J.G.B. performed the functional cAMP and reporter gene assays. R.M. was involved in data collection and processing. A.G.W.L. processed the final data, solved and refined the structure, and assisted with manuscript preparation. The overall project management and manuscript preparation were by R.H., C.G.T. and G.F.X.S.

Author Information Co-ordinates and structure factors have been submitted to the PDB database under accession code 2vt4. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the paper on www.nature.com/nature. Correspondence and requests for materials should be addressed to C.G.T. (cgt@mrc-lmb.cam.ac.uk) or G.F.X.S. (gfx@mrc-lmb.cam.ac.uk).

METHODS

Purification and crystallization. Baculovirus expression in High 5 cells, membrane preparation, solubilization, IMAC and alprenolol sepharose chromatography were all performed as described previously[42], except that solubilization and IMAC were performed in buffers containing the detergent decylmaltoside and the detergent was exchanged on the alprenolol sepharose column to octylthioglucoside; purified receptor was eluted from the alprenolol sepharose with cyanopindolol (30 μM). The buffer was exchanged to 10 mM Tris-HCl, pH 7.7, 50 mM NaCl, 0.1 mM EDTA, 0.35% octylthioglucoside and 0.5 mM cyanopindolol during concentration to give a final receptor concentration of 5.5–6.0 mg ml^{-1}.

Using the thermally stabilized protein, a wide crystal screen was performed in four different detergents. A total of 58 mg of receptor was used to set up 17,800 crystallization trials in an MRC ultraviolet transparent crystallization plate and imaged with the MRC multiwavelength imaging system at 380 nm. Promising looking crystals were then imaged at 280 nm to exclude salt and detergent crystals. The receptor crystallization in octylthioglucoside was optimised by vapour diffusion at 18 °C with hanging drops after addition of an equal volume of reservoir solution (0.1 M ADA, pH 6.9–7.3, and 29–32% PEG 600). Crystals were mounted on Hampton CrystalCap HT loops and were cryo-cooled in liquid nitrogen. Cryoprotection of crystals was achieved by increasing the PEG 600 concentration in the drop to 55–70%.

Data collection, structure solution and refinement. The first diffraction patterns from microcrystals grown in the primary crystallization screens were tested with a 5 μm beam on beamline ID13 (ref. 43) at the European Synchrotron Radiation Facility, Grenoble. The best crystallization conditions were refined to improve diffraction quality and the optimized crystals were then screened at ID23-2 with a 10 μm focused beam; the micro-beams helped to deal with heterogeneous diffraction within a single crystal. Diffraction data were collected with a Mar 225 CCD detector on the microfocus beamline ID23-2 (wavelength, 0.8726 Å) using three positions on a single cryo-cooled crystal (100 K) with dimensions 240 × 40 × 10 μm.

During the refinement of the model with PHENIX, tight non-crystallographic symmetry restraints ($\sigma = 0.025$ Å) were applied to chains A and D and to chains B and C, with an accompanying reduction in R_{free}. Molecules A and D differ from molecules B and C by 0.46 Å r.m.s.d. on main-chain atoms excluding the N terminus of kinked H1. Molecules B and C have 0.18 Å r.m.s.d. for 272 residues. Molecules A and D have 0.22 Å r.m.s.d. for 272 residues. The cyanopindolol ligand, detergent, water molecules and sodium ions were added at a late stage of refinement. Non-crystallographic restraints were not applied to detergent and water molecules. The correct side-chain rotamer of Ser 211 was ambiguous with both *gauche*$^+$ and *trans* rotamers giving an equally good fit to the electron density after refinement. The *gauche*$^+$ conformation was chosen because the *trans* conformation resulted in a short contact of 2.8 Å between the β-carbon of Ser 211 and the carbonyl oxygen of residue 207. In addition, the *gauche*$^+$ conformation allows the serine hydroxyl to act as both a hydrogen bond donor and an acceptor, whereas it can only act as an acceptor in the *trans* conformation. An overview of the B-factor distribution for β_1AR molecules A and B is shown in Supplementary Fig. 6, alongside rhodopsin (PDB code 1GZM) and β_2AR-T4 (PDB code 2RH1).

G-protein-coupling assays and cAMP measurement. Stable cell lines expressing the βAR-m23 mutation were made. A stable clonal CHO-K1 cell line expressing a *CRE-SPAP* reporter gene (six cAMP response elements (*CRE*) upstream of a secreted placental alkaline phosphatase (*SPAP*) gene) was transfected with plasmid pcDNA3 containing the βAR-m23 complementary DNA. The transfected cells were selected by neomycin resistance (1 mg ml^{-1}; for the turkey receptor) and hygromycin resistance (200 μg ml^{-1}; for the *CRE-SPAP* reporter gene) for three weeks, and then single clones were isolated by dilution cloning to give clonal lines (CHO-m23-SPAP cells).

Whole-cell binding assays using ^{3}H-CGP12177 were performed and the K_i values of available agonists and antagonists were determined from competition curves (J.G.B., unpublished observation). The ability of βAR-m23 to couple to G proteins was assessed by using a *CRE-SPAP* reporter assay as described previously[50]. In brief, confluent cells in serum-free medium were incubated for 5 h with the agonist, or after pre-treatment with antagonist for 1 h. Cells were then incubated for a further hour in the absence of ligands and the level of secreted alkaline phosphates then determined by a colorimetric reaction using pNPP. To determine what conformation βAR-m23 has in the absence of ligand, the effect of a known inverse agonist ICI 118551 was tested on the CHO-m23-SPAP cells. Confluent cells were pre-labelled with ^{3}H-adenine in serum-free medium for 2 h, were removed and then the cells were incubated in 100 μM isobutylmethylxanthine (IBMX, a non-specific phosphodiesterase inhibitor) and ICI 118551. After 5 h the reaction was terminated and ^{3}H-cAMP separated from other ^{3}H-nucleotides by sequential Dowex and alumina column chromatography, as described previously[51]. Under these conditions any inverse agonist affects of ICI 118551 would have been seen[52]. Data for all experiments were analysed by GraphPad Prism; all data are presented as mean ± s.e.m. of triplicate determinations, where n is the number of separate experiments. Using the same assay, cyanopindolol appeared to be a very weak partial agonist.

All agonist ligands used were examined in the parent CHO-SPAP cells (that is, cells expressing the reporter but not βAR-m23); no effects were seen in response to any of the ligands over a 10^7 concentration range despite an increase in CRE-SPAP production in response to 3 μM forskolin. This suggests that all the responses in CHO-m23-SPAP cells were indeed occurring by means of βAR-m23.

Figure production. The alignment of receptors (Supplementary Fig. 1) was performed using ClustalW (MacVector), and Supplementary Figs 4–6 and 8–10 were made using PyMol (DeLano Scientific LLC).

50. Baker, J. G., Hall, I. P. & Hill, S. J. Agonist actions of "beta-blockers" provide evidence for two agonist activation sites or conformations of the human β_1-adrenoceptor. *Mol. Pharmacol.* **63**, 1312–1321 (2003).
51. Donaldson, J., Brown, A. M. & Hill, S. J. Influence of rolipram on the cyclic 3′,5′-adenosine monophosphate response to histamine and adenosine in slices of guinea-pig cerebral cortex. *Biochem. Pharmacol.* **37**, 715–723 (1988).
52. Baker, J. G., Hall, I. P. & Hill, S. J. Agonist and inverse agonist actions of beta-blockers at the human β_2-adrenoceptor provide evidence for agonist-directed signaling. *Mol. Pharmacol.* **64**, 1357–1369 (2003).

Vol 454|24 July 2008|doi:10.1038/nature07114

LETTERS

The characteristic blue spectra of accretion disks in quasars as uncovered in the infrared

Makoto Kishimoto[1,2], Robert Antonucci[3], Omer Blaes[3], Andy Lawrence[2], Catherine Boisson[4], Marcus Albrecht[5] & Christian Leipski[3]

Quasars are thought to be powered by supermassive black holes accreting surrounding gas[1–3]. Central to this picture is a putative accretion disk which is believed to be the source of the majority of the radiative output[2–4]. It is well known, however, that the most extensively studied disk model[5]—an optically thick disk which is heated locally by the dissipation of gravitational binding energy—is apparently contradicted by observations in a few major respects[6,7]. In particular, the model predicts a specific blue spectral shape asymptotically from the visible to the near-infrared[5,8], but this is not generally seen in the visible wavelength region where the disk spectrum is observable[9–13]. A crucial difficulty has been that, towards the infrared, the disk spectrum starts to be hidden under strong, hot dust emission from much larger but hitherto unresolved scales, and thus has essentially been impossible to observe. Here we report observations of polarized light interior to the dust-emitting region that enable us to uncover this near-infrared disk spectrum in several quasars. The revealed spectra show that the near-infrared disk spectrum is indeed as blue as predicted. This indicates that, at least for the outer near-infrared-emitting radii, the standard picture of the locally heated disk is approximately correct.

A success of the most extensively studied disk model is that it gives the radiative output peak approximately correctly in the ultraviolet ($\sim$0.01–0.4 μm) wavelengths for the case of a supermassive black hole. This is observed for a generic spectral energy distribution[14] of quasars, the most luminous examples of active galactic nuclei (AGNs). However, it has long been known that the model apparently shows a few major contradictions with observations[6,7]. One of the disagreements, and perhaps the most easily comprehensible one, is the spectral shape of the radiation. From the basic hypothesis of the model, the effective disk temperature T is fixed as a function of radius r as $T \propto r^{-3/4}$ over a broad range of radii. This leads to a well-known blue spectral-shape limit, $F_\nu \propto \nu^{+1/3}$ (where F_ν is flux per frequency ν), being asymptotically reached at long wavelengths, from the visible (also called the optical; $\sim$0.4–1 μm) to the near-infrared ($\sim$1–2 μm) for AGN disks. In contrast, many studies have shown that the general AGN spectral shape observed at optical–ultraviolet wavelengths is much redder, with a spectral slope α (where $F_\nu \propto \nu^{\alpha}$) between -0.2 and -1, and is never as blue as this spectral shape[9–13].

The predicted blue spectral-shape limit is strictly true in the simplest assumption of local black-body emission. In more sophisticated disk atmosphere models[8], the spectrum generally becomes slightly redder at optical wavelengths, owing to various opacity effects and deviation from local thermodynamic equilibrium, but discrepancies between the model and observed spectra still remain[15]. However, the redder model slopes at optical wavelengths form a wider concave spectrum that shifts the bluer limit above to longer wavelengths, into the near-infrared. The observed spectra certainly appear to become bluer from the short-ultraviolet wavelengths to the optical. The crucial observational difficulty here has been that the disk spectrum starts to be hidden under the hot dust thermal emission that begins at wavelengths greater than $\sim$1 μm, a limit set by the sublimation temperature of dust grains ($\sim$1,500 K). These dust grains exist at larger spatial scales, in a configuration often thought to have a torus-like geometry but which is generally not yet spatially resolvable. Therefore, it has been virtually impossible to observe the underlying near-infrared disk spectrum[16]. We note that, contrary to early spectral-fitting studies[3,4,16,17], the infrared component is no longer thought to be non-thermal, and thus cannot be extrapolated to underlie the optical spectrum. In the early studies, doing so effectively made the inferred disk spectrum on top of the non-thermal spectrum bluer.

We argue here that this hidden part of the disk spectrum can be revealed by observing the near-infrared polarized light. Optical continua of many directly visible AGNs called Type 1s (Seyfert 1 galaxies and quasars) are known to be linearly polarized at a level $\lesssim$1%. The polarization position angle in these Type-1 cases is mostly parallel to the rotation axis of the putative accretion disk, where the axis can be probed by the linear jet-like structure of radio emission[18–20]. (This is in contrast to the cases in hidden AGNs, or Type 2s, which show high polarization at perpendicular position angles and strong, broad lines in polarized light, and which are not the subject of the present paper.) This polarization in Type 1s is interpreted as an indication of an equatorial scattering region that is optically thin and surrounds the disk. In many Seyfert 1 galaxies, broad emission lines are polarized at a much lower level than the continuum polarization, and at different position angles[19,20], indicating that the scatterers reside roughly at the same spatial scales as the broad-line-emitting clouds.

At least in several quasars, the emission line polarization vanishes—the spectrum of optical polarized light shows very little or no emission line flux[21,22]. This is very likely to indicate that the scatterers reside interior to the broad-line-emitting clouds. In these Type-1 cases, the scatterers are thought to be electrons rather than dust grains, as the scattering region is interior to the dust sublimation radius. Because electron scattering is wavelength independent, the polarized light therefore produces a copy of the spectrum originating in the region interior to the scattering region. (We note that this electron scattering is conceptually different from that discussed in previous works[23,24], which was assumed to be intrinsic to the accretion disk atmosphere and gave rise to the prediction of perpendicular position angles.) In the optical polarized light from these quasars, which excludes the emission from the broad-line region, we actually

[1]Max-Planck-Institut für Radioastronomie, Auf dem Hügel 69, 53121 Bonn, Germany. [2]Scottish Universities Physics Alliance, Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK. [3]Physics Department, University of California, Santa Barbara, California 93106, USA. [4]LUTH, FRE 2462 du CNRS, associée à l'Université Denis Diderot, Observatoire de Paris, Section de Meudon, F–92195 Meudon Cedex, France. [5]Instituto de Astronomía, Universidad Católica del Norte, Avenida Angamos 0610, Antofagasta 1270709, Chile.

found a hydrogen Balmer-edge feature in absorption, which we believe originates in the disk and specifically indicates that the emission is thermal and optically thick[21,22].

We can then use the same polarized light, but in the near-infrared, to reveal the hidden spectrum of the disk, by removing the dust radiation from the torus exterior to the scattering region. In previous work[25], we suggested that this method appears to work in at least one quasar that has polarization data with high signal-to-noise ratio at two wavelength bands in the near-infrared. However, crucial information needed at the time was whether the behaviour of near-infrared polarized light is consistent and systematic in different objects. If it is, this would critically argue against, for example, a possible secondary polarization component related to dust grains arising in the near-infrared. Therefore, we undertook the near-infrared polarimetry of five other quasars. The targets were selected to be polarized in their optical continua but essentially not in their emission lines, to ensure that the scattering was interior to the broad-line region. These properties of the quasars were either already known[22] or were determined in our optical polarimetric survey and follow-up spectropolarimetry.

In Fig. 1 we show the spectra of linearly polarized light measured in the near-infrared broad-band imaging polarimetry, and optical spectropolarimetry, of the six quasars (including the one studied in the previous work). Details of the measurements, as well as the procedures for removing instrumental polarizations, can be found in the Supplementary Information. Generally, polarization degrees observed for these quasars are ~1% at ~0.5 μm, gradually decreasing to ~0.5% at ~2 μm, and position angles are essentially constant over the near-ultraviolet–optical–near-infrared wavelengths, for a given object. A significant result here is that all the objects behave in a similar and systematic way, showing blue polarized-light spectra. Whereas the total-light spectra begin to increase in νF_ν at around 1 μm as wavelength increases, owing to the onset of dust emission, all the polarized-light spectra, which eliminate dust, display a rapid decrease in νF_ν, with a shape of approximately power-law form.

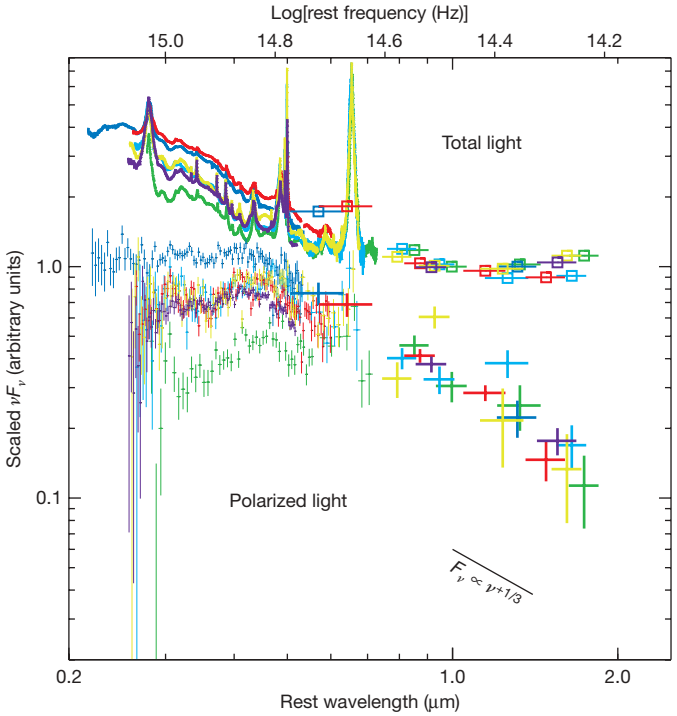


Figure 1 | Overlay of the polarized- and total-light spectra observed in six different quasars. We plot scaled νF_ν data: Q0144-3938 (redshift $z = 0.244$), green; 3C 95 ($z = 0.616$), blue; CTS A09.36 ($z = 0.310$), light blue; 4C 09.72 ($z = 0.433$), red; PKS 2310-322 ($z = 0.337$), yellow. Plotted in purple are the data for Ton 202 ($z = 0.366$) from a previous paper[25]. Total-light spectra, shown as bold traces in the optical and as squares in the near-infrared, are normalized at 1 μm in the rest frame, by interpolation (except that of 3C 95, which we normalized by νF_ν observed at 1.3 μm in the rest frame). Polarized-light spectra, shown as light points in the optical and as bold points in the near-infrared (vertical error bars, 1σ), are separately normalized, also at 1 μm, by fitting a power law to the near-infrared polarized-light spectra. For both total-light and polarized-light data, horizontal bar lengths indicate bandwidth. The normalized polarized-light spectra are arbitrarily shifted downwards by a factor of three relative to the normalized total-light spectra, for clarity. The total-light spectra begin to increase in νF_ν at wavelengths around, or slightly greater than, 1 μm. In contrast, the polarized-light spectra all consistently and systematically decrease in νF_ν towards long wavelengths, showing a blue shape of approximately power-law form.

The measurement of the spectral index α in F_ν (such that $F_\nu \propto \nu^\alpha$) for each object is shown in Fig. 2. The measured slopes are consistent with each other within their errors, and the individual slopes as well as their average clearly point to a spectral shape much bluer than those observed in the ultraviolet–optical. Surprisingly, they are all consistent with the $F_\nu \propto \nu^{+1/3}$ shape. The weighted mean of the measured slopes is $\alpha = +0.44 \pm 0.11$. Although the sample size is small, there does not appear to be any luminosity dependence, as seen in Fig. 2, and we did not find dependencies on black-hole masses $M_{\rm BH}$ or Eddington ratios $L/L_{\rm Edd}$ derived from the width of the Hβ Balmer line. This is expected if the near-infrared spectrum is in the long-wavelength limit of the disk model, which is independent of parameters such as black-hole mass and Eddington ratio. In this case, by regarding each measurement as a measurement of the same quantity, the weighted mean over the measurements becomes physically meaningful. We note that if we formally convert the mean slope to the radial temperature distribution for the case of an optically thick disk, we obtain $T \propto r^{-0.78 \pm 0.03}$, consistent with the predicted dependence $T \propto r^{-3/4}$.

The systematic behaviour of the near-infrared polarized light, as well as the constancy of the position angles over all wavelengths, strongly argues against there being any secondary polarization contamination.

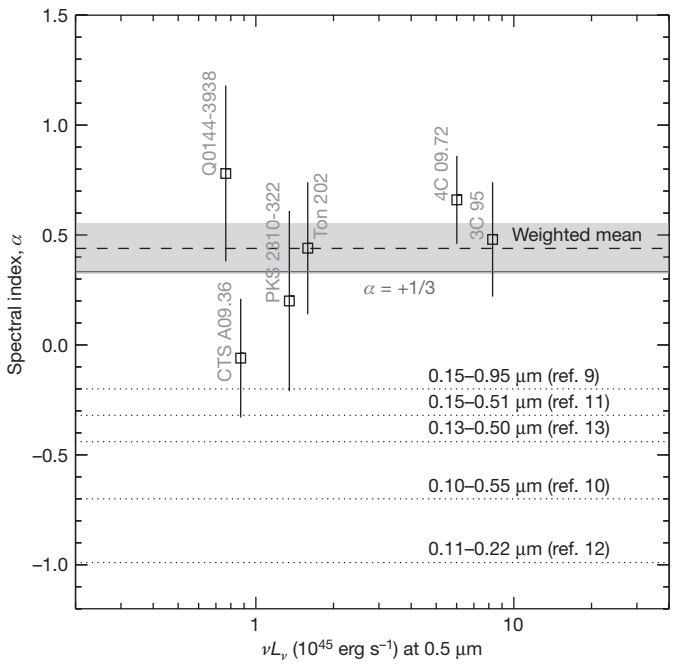


Figure 2 | Spectral index of polarized light spectra. We plot α (in $F_\nu \propto \nu^\alpha$) against νL_ν for total light at 0.51 μm. The index was measured using a power-law fit for each near-infrared polarized-light spectrum (note the different wavelength ranges covered depending on the redshift) and is shown with 1σ error bars. A weighted mean of the spectral index measurements is shown dashed; the shaded area represents its deduced 1σ uncertainty. The mean or median slopes of the ultraviolet–optical total-light spectra derived in various other studies[9–13] are also shown.

We might worry that the polarized light would be affected if the corresponding spatial scale of the disk emission at long wavelengths were to become large and finite in comparison with the size of the scattering region. However, this seems unlikely, because the half-light radius of the disk (within which half of the total light is emitted) even at 2 μm is still much smaller ($\sim 400R_S$, where $R_S = 2GM_{BH}/c^2$) than at least the radius of the broad-line region[26] ($\sim 4000R_S$), in the case of an untruncated multi-temperature black-body disk for our quasars. Significant geometrical effects will not occur unless the disk emission size becomes almost the same as the scattering region size. Therefore, the spectra of near-infrared polarized light are very likely to reveal the intrinsic spectra of accretion disks.

The measured slopes, being as blue as the slope of the predicted spectral shape $F_\nu \propto \nu^{+1/3}$, strongly suggest that, at least in the outer near-infrared-emitting radii, the standard but hitherto unverified picture of the disk being optically thick and locally heated is approximately correct. In this case, an implication is that other model problems at shorter wavelengths are associated with, or originate from, our lack of understanding of the inner regions of the same disks. We note that disk irradiation in limiting cases can contribute to the heating without changing the $\nu^{+1/3}$ spectral shape at long wavelengths; but it would not dominate the local internal heating in the outer radii considered here, and thus is not directly relevant except in some very specific cases[27,28].

The standard disk is also well known to be gravitationally unstable at large radii[29]. These radii may correspond to those emitting in the infrared[30] ($\sim 800R_S$ for our quasars). In this case, if the disk is truncated at such a radius, the spectrum will show a break, becoming even bluer at the longest wavelengths[25]. Although statistically insignificant, our data do suggest that the near-infrared slope is slightly bluer than the spectral shape $F_\nu \propto \nu^{+1/3}$, with a hint of possibly becoming bluer at longer wavelengths. This can be followed up by extending the wavelength coverage with similar polarized Type-1 AGNs at lower redshifts. Such future measurements may show the way to probe how and where the disk ends and how material is being supplied to the nucleus.

Received 30 January; accepted 15 May 2008.

1. Salpeter, E. E. Accretion of interstellar matter by massive objects. *Astrophys. J.* **140,** 796–800 (1964).
2. Lynden-Bell, D. Galactic nuclei as collapsed old quasars. *Nature* **223,** 690–694 (1969).
3. Shields, G. A. Thermal continuum from accretion disks in quasars. *Nature* **272,** 706–708 (1978).
4. Malkan, M. A. The ultraviolet excess of luminous quasars. II - Evidence for massive accretion disks. *Astrophys. J.* **268,** 582–590 (1983).
5. Shakura, N. I. & Sunyaev, R. A. Black holes in binary systems. Observational appearance. *Astron. Astrophys.* **24,** 337–355 (1973).
6. Antonucci, R. in *High Energy Processes in Accreting Black Holes (Astron. Soc. Pacif. Conf. Ser. 161)* (eds Poutanen, J. & Svensson, R.) 193–203 (Astronomical Society of the Pacific, San Francisco, 1999).
7. Koratkar, A. & Blaes, O. The ultraviolet and optical continuum emission in active galactic nuclei: the status of accretion disks. *Publ. Astron. Soc. Pacif.* **111,** 1–30 (1999).
8. Hubeny, I., Agol, E., Blaes, O. & Krolik, J. H. Non-LTE models and theoretical spectra of accretion disks in active galactic nuclei. III. Integrated spectra for hydrogen-helium disks. *Astrophys. J.* **533,** 710–728 (2000).
9. Neugebauer, G. *et al.* Continuum energy distributions of quasars in the Palomar-Green Survey. *Astrophys. J. Suppl. Ser.* **63,** 615–644 (1987).
10. Cristiani, S. & Vio, R. The composite spectrum of quasars. *Astron. Astrophys.* **227,** 385–393 (1990).
11. Francis, P. J. *et al.* A high signal-to-noise ratio composite quasar spectrum. *Astrophys. J.* **373,** 465–470 (1991).
12. Zheng, W., Kriss, G. A., Telfer, R. C., Grimes, J. P. & Davidsen, A. F. A. Composite HST spectrum of quasars. *Astrophys. J.* **475,** 469–479 (1997).
13. Vanden Berk, D. E. *et al.* Composite quasar spectra from the Sloan Digital Sky Survey. *Astron. J.* **122,** 549–564 (2001).
14. Sanders, D. B., Phinney, E. S., Neugebauer, G., Soifer, B. T. & Matthews, K. Continuum energy distribution of quasars - Shapes and origins. *Astrophys. J.* **347,** 29–51 (1989).
15. Davis, S. W., Woo, J.-H. & Blaes, O. M. The UV continuum of quasars: Models and SDSS spectral slopes. *Astrophys. J.* **668,** 682–698 (2007).
16. Malkan, M. in *Theory of Accretion Disks (NATO ASIC Proc. 290)* (ed. Meyer, F.) 19–28 (Kluwer Academic, Norwell, Massachusetts, 1989).
17. Malkan, M. A. & Filippenko, A. V. The stellar and nonstellar continua of Seyfert galaxies: Nonthermal emission in the near-infrared. *Astrophys. J.* **275,** 477–492 (1983).
18. Antonucci, R. R. J. Optical polarization position angle versus radio structure axis in Seyfert galaxies. *Nature* **303,** 158–159 (1983).
19. Smith, J. E. *et al.* A spectropolarimetric atlas of Seyfert 1 galaxies. *Mon. Not. R. Astron. Soc.* **335,** 773–798 (2002).
20. Smith, J. E. *et al.* Seyferts on the edge: polar scattering and orientation-dependent polarization in Seyfert 1 nuclei. *Mon. Not. R. Astron. Soc.* **350,** 140–160 (2004).
21. Kishimoto, M., Antonucci, R. & Blaes, O. A first close look at the Balmer-edge behaviour of the quasar big blue bump. *Mon. Not. R. Astron. Soc.* **345,** 253–260 (2003).
22. Kishimoto, M., Antonucci, R., Boisson, C. & Blaes, O. The buried Balmer edge signatures from quasars. *Mon. Not. R. Astron. Soc.* **354,** 1065–1092 (2004).
23. Webb, W., Malkan, M., Schmidt, G. & Impey, C. The wavelength dependence of polarization of active galaxies and quasars. *Astrophys. J.* **419,** 494–514 (1993).
24. Impey, C. D., Malkan, M. A., Webb, W. & Petry, C. E. Ultraviolet spectropolarimetry of high-redshift quasars with the Hubble Space Telescope. *Astrophys. J.* **440,** 80–90 (1995).
25. Kishimoto, M., Antonucci, R. & Blaes, O. The dust-eliminated shape of quasar spectra in the near-infrared: a hidden part of the big blue bump. *Mon. Not. R. Astron. Soc.* **364,** 640–648 (2005).
26. Bentz, M. C., Peterson, B. M., Pogge, R. W., Vestergaard, M. & Onken, C. A. The radius-luminosity relationship for active galactic nuclei: the effect of host-galaxy starlight on luminosity measurements. *Astrophys. J.* **644,** 133–142 (2006).
27. Blaes, O. M. in *Accretion Discs, Jets and High Energy Phenomena in Astrophysics* (eds Beskin, V. *et al.*) 137–185 (Springer, Berlin, 2003).
28. Agol, E. & Krolik, J. H. Magnetic stress at the marginally stable orbit: Altered disk structure, radiation, and black hole spin evolution. *Astrophys. J.* **528,** 161–170 (2000).
29. Shlosman, I. & Begelman, M. C. Self-gravitating accretion disks in active galactic nuclei. *Nature* **329,** 810–812 (1987).
30. Goodman, J. Self-gravity and quasi-stellar object discs. *Mon. Not. R. Astron. Soc.* **339,** 937–948 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The UK Infrared Telescope (UKIRT) is operated by the Joint Astronomy Centre on behalf of the Science and Technology Facilities Council of the UK. We thank the Department of Physical Sciences, University of Hertfordshire, for providing the IRPOL2 polarimetry facility for the UKIRT. This research is partially based on observations collected at the European Southern Observatory, Chile.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to M.K. (mk@mpifr-bonn.mpg.de).

LETTERS

Medium-scale carbon nanotube thin-film integrated circuits on flexible plastic substrates

Qing Cao[1], Hoon-sik Kim[2], Ninad Pimparkar[7], Jaydeep P. Kulkarni[7], Congjun Wang[2], Moonsub Shim[2], Kaushik Roy[7], Muhammad A. Alam[7] & John A. Rogers[1–6]

The ability to form integrated circuits on flexible sheets of plastic enables attributes (for example conformal and flexible formats and lightweight and shock resistant construction) in electronic devices that are difficult or impossible to achieve with technologies that use semiconductor wafers or glass plates as substrates[1]. Organic small-molecule and polymer-based materials represent the most widely explored types of semiconductors for such flexible circuitry[2]. Although these materials and those that use films or nanostructures of inorganics have promise for certain applications, existing demonstrations of them in circuits on plastic indicate modest performance characteristics that might restrict the application possibilities. Here we report implementations of a comparatively high-performance carbon-based semiconductor consisting of sub-monolayer, random networks of single-walled carbon nanotubes to yield small- to medium-scale integrated digital circuits, composed of up to nearly 100 transistors on plastic substrates. Transistors in these integrated circuits have excellent properties: mobilities as high as 80 $cm^2 V^{-1} s^{-1}$, subthreshold slopes as low as 140 m V dec^{-1}, operating voltages less than 5 V together with deterministic control over the threshold voltages, on/off ratios as high as 10^5, switching speeds in the kilohertz range even for coarse (~100-μm) device geometries, and good mechanical flexibility—all with levels of uniformity and reproducibility that enable high-yield fabrication of integrated circuits. Theoretical calculations, in contexts ranging from heterogeneous percolative transport through the networks to compact models for the transistors to circuit level simulations, provide quantitative and predictive understanding of these systems. Taken together, these results suggest that sub-monolayer films of single-walled carbon nanotubes are attractive materials for flexible integrated circuits, with many potential areas of application in consumer and other areas of electronics.

Efforts to develop polymer and small-molecule semiconductors for electronics have yielded several impressive demonstrations, including integrated circuits with more than 1000 transistors[3], flexible displays[3,4], sensor sheets[5] and other systems[6,7]. In all cases, however, the field-effect mobilities of the transistors are modest: typically ~1 $cm^2 V^{-1} s^{-1}$ for isolated devices[8,9] and <0.05 $cm^2 V^{-1} s^{-1}$ in integrated circuits[3–7]. Although these properties are sufficient for electrophoretic displays and certain other applications, improvements in the materials would expand the possibilities[1]. Separately, for any given application, increases in mobility relax the requirements on critical feature sizes in the circuits (for example transistor channel lengths) and tolerances on their multilevel registration, which can be exploited to reduce the cost of the plastic substrates and patterning systems to achieve roll-to-roll fabrication by dry printing[10] or ink-jet printing[11].

Recently developed carbon-based semiconducting nanomaterials, especially single-walled carbon nanotubes (SWNTs), might provide an opportunity to achieve extremely high intrinsic mobilities, high current-carrying capacities and exceptional mechanical/optical characteristics, in bendable formats on plastic substrates[12]. Although isolated SWNTs are not relevant to the applications contemplated here, recent work shows that sub-monolayer random networks[13–16] or aligned arrays[17,18] of SWNTs can serve as thin-film semiconductors which, in the best cases, inherit the exceptional properties of the tubes, for example device mobilities up to ~2,500 $cm^2 V^{-1} s^{-1}$, on-state currents above several milliamperes, and cut-off frequencies above 1 GHz for devices on plastic. The network geometry is of particular interest for flexible electronics because it can be easily achieved by printing SWNTs from solution suspensions[19]. The present work demonstrates implementations of SWNT networks in flexible integrated circuits on plastic that have attractive characteristics, together with corresponding theoretical models and simulation tools that capture all of the key aspects.

The system layouts (Fig. 1a) exploit architectures similar to those in established silicon integrated circuits. A thin (50-μm) sheet of polyimide serves as the substrate. Random networks of SWNTs grown by chemical vapour deposition and subsequently transfer printed onto the polyimide form the semiconductor layer[17]. Source and drain (S–D) electrodes of gold serve as low-resistance contacts to these networks, as determined by scaling studies (Supplementary Fig. 1). Although roughly one-third of the SWNTs are metallic, purely metallic transport pathways between the S–D electrodes can be eliminated by suitably engineering the average tube lengths and the network layouts: for the present purposes, we used soft lithography and reactive-ion etching to cut fine lines into the networks. The resulting network strips are oriented along the overall direction of transport, with widths designed to reduce the probability of metallic pathways below a practical level without significantly reducing the effective thin-film mobility of the network.

Figure 1b shows a scanning electron micrograph of a region of an integrated circuit just before deposition of the gate dielectric. A magnified view of a part of the SWNT network in the channel of one of the devices (Fig. 1c; the S–D electrodes are to the right and left, outside the field of view) reveals narrow, dark horizontal lines, corresponding to the etched regions. The critically important role of these features in determining the electrical characteristics can be quantified through first-principles modelling studies that consider percolative transport through sticks with average lengths and layouts (for example etched lines, densities of SWNTs and so on) corresponding to experiment[20]. Fig. 1d shows the distribution of current flow in a typical case, in which the colour indicates the current density in the

[1]Department of Chemistry, [2]Department of Materials Science and Engineering, [3]Department of Electrical and Computer Engineering, [4]Department of Mechanical Science and Engineering, [5]Frederick-Seitz Materials Research Laboratory, [6]Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA. [7]School of Electrical and Computer Engineering, Network for Computational Nanotechnology, Purdue University, West Lafayette, Indiana 47907, USA.

on-state of the device. In addition to providing guidance on optimal design (Fig. 2a), these simulations reveal that networks with this geometry and coverage (~0.6%) distribute current evenly, thereby serving as an effective film for transport. A typical device incorporates ~16,000 individual SWNTs. The circuits are completed in top-gate configurations by deposition and patterning of high-capacitance, hysteresis-free dielectrics enabled by low operating voltages (~40 nm of hafnium dioxide) directly on the tubes, followed by gate metallization and the addition of vias and interconnects. Figure 1e shows a representative system, complete with arrays of isolated enhancement-mode (lower right region) and depletion-mode (lower middle region) transistors, various logic gates (lower left part), and two four-bit row decoders each twenty logic gates in size (middle and upper parts). Fabrication details are further described in the Methods.

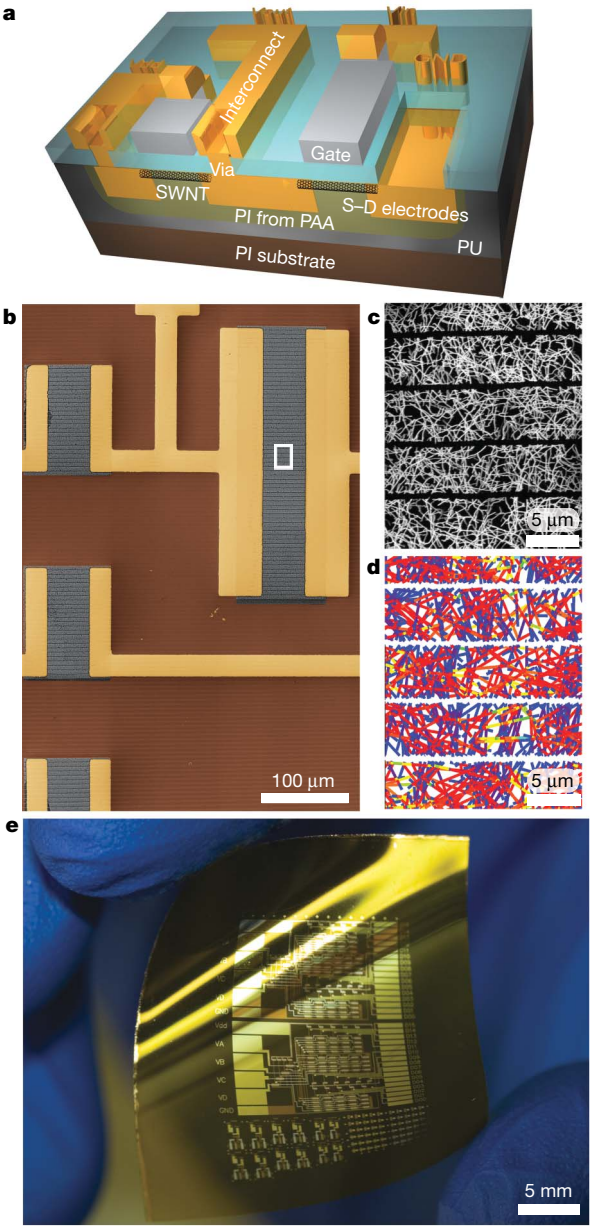


Figure 1 | Illustration, scanning electron microscope images, theoretical modelling results and photographs of flexible SWNT integrated circuits on plastic. a, Cross-sectional diagram of a SWNT PMOS inverter on a PI substrate. PI, polyimide; PU, polyurethane; PAA, polyamic acid; $V_{dd} \equiv$ Vdd, common power supply voltage; $V_{out} \equiv$ Vout, output voltage; $V_{in} \equiv$ Vin, input voltage; GND, common ground. **b**, Scanning electron microscope image of part of the SWNT circuit, made before deposition of the gate dielectric, gate or gate-level interconnects. The S–D electrodes (gold) and substrates (brown) had been colourized to highlight the SWNT network strips (black and grey) that form the semiconductor. **c**, Magnified view of the network strips corresponding to a region of the device channel highlighted with the white box in **b**. **d**, Theoretical modelling results for the normalized current distribution in the on-state of the device (view as in **c**), where colour indicates current density (yellow, high; red, medium; blue, low). **e**, Photograph of a collection of SWNT transistors and circuits on a thin sheet of plastic (PI).

Figure 2 summarizes measurements on individual transistors. Figure 2a illustrates the predicted and measured influences of the geometry of the etched lines described above on devices with coarse dimensions (that is, channel lengths $L_C = 100\ \mu m$), selected to be compatible with established low-cost patterning techniques such as screen printing[21], and with sufficiently high densities of SWNTs to achieve good performance and uniformity as a thin-film semiconductor. For widths of ~5 μm, the etched lines increase the on/off ratios by up to four orders of magnitude, while reducing the transconductances (g_m) by only ~40%. Figure 2b, c shows characteristics of transistors with this geometry, illustrating well-behaved responses with minimal hysteresis and with excellent channel-width-normalized transconductances (as high as 0.15 $\mu S\ \mu m^{-1}$ and typically 0.12 $\mu S\ \mu m^{-1}$ for $L_C \geq 50\ \mu m$, which corresponds to an estimated cut-off frequency of >100 kHz), device mobilities (μ_{eff}; as high as ~80 $cm^2\ V^{-1}\ s^{-1}$ and typically ~70 $cm^2\ V^{-1}\ s^{-1}$ as calculated using standard models of metal–oxide–semiconductor field-effect transistors with measured gate capacitances (Supplementary Fig. 2), for both the linear and the saturation regimes), and subthreshold swings (S; as low as 140 mV dec^{-1} and typically ~200 mV dec^{-1}).

The transconductances and the subthreshold behaviours, in particular, exceed those that have been demonstrated in flexible integrated circuits on plastic with organic thin-film semiconductors ($g_m < 0.02\ \mu S\ \mu m^{-1}$ for $L_C \approx 50\ \mu m$, $S > 140$ mV dec^{-1})[22,23] or with silicon nanowires ($g_m < 0.01\ \mu S\ \mu m^{-1}$ for $L_C \approx 50\ \mu m$, $S > 280$ mV dec^{-1})[24], and are competitive with the best reports of p-channel single-crystalline silicon ribbons ($g_m \approx 0.25\ \mu S\ \mu m^{-1}$ for $L_C = 50\ \mu m$, $S \approx 230$ mV dec^{-1})[25]. Under low-to-moderate bias conditions, the on/off ratios can be as high as 10^5 (Fig. 2f and Supplementary Fig. 3a), and typically ~10^3, for transistors with this geometry. The inset in Fig. 2b and Supplementary Fig. 4a show a decrease in the on/off ratio with increasing drain–source voltage (V_{DS}), which is due primarily to the slightly ambipolar nature of the device operation. These ratios also decrease with L_C (Supplementary Fig. 1b). The favourable d.c. properties of long-channel devices can be achieved at short L_Cs, for improved operating speeds, either by use of correspondingly shorter SWNTs and narrower etched stripes, as suggested by modelling results, or by using pre-enriched semiconducting SWNTs[26].

The threshold voltage (V_T) can be controlled by using gate metals with different work functions, because the high-capacitance gate dielectrics reduce the relative contribution of voltage across the dielectric to V_T (ref. 27). For example, replacing gold with aluminium as the gate metal shifts V_T by −(0.6–0.8) V, thereby changing the device operation from depletion mode to enhancement mode (Fig. 2b). Systematic bending tests of individual devices and inverters showed no significant change in device performance during inward or outward bending to radii as small as ~5 mm (Fig. 2d). Collectively, these properties are as good as or better than those of previously reported devices based on SWNT random networks, in spite of the moderate decreases in g_m associated with the etching procedures. Transistors that use dense, perfectly aligned arrays of SWNTs have improved performance, that is, device mobility up to 2,500 $cm^2\ V^{-1}\ s^{-1}$, but these layouts cannot be formed readily with solution deposition techniques[17]. As such, they are not relevant for the type of printed, flexible electronics applications contemplated here.

For use in integrated circuits, the yields and performance uniformity of the transistors are critically important. We examined these aspects through measurements on more than 100 devices (Fig. 2e, f and Supplementary Fig. 5). The results show standard deviations of ~20% for the normalized on-state current (I_{on}) and ~0.05 V for V_T. The former result is quantitatively in agreement with percolation theory, illustrated also in Fig. 2e. Although on/off ratios vary by roughly two orders of magnitude, most of the values are $>10^3$. The distribution (Fig. 2f) indicates no correlation with V_T (suggesting the importance of extrinsic doping effects on SWNTs[28]), and has a width which is much larger than that predicted by percolation models (Fig. 2f, inset) that do not explicitly include effects of S–D contacts. These results strongly indicate that the variation in on/off ratio results from electron conduction caused by tunnelling through the Schottky barriers at the S–D contacts (Fig. 2f, inset)[29]. Although unnecessary for the circuits reported here, doping techniques similar to those demonstrated in single-SWNT devices can be used to suppress the ambipolar behaviour and improve on/off ratio uniformity[30]. Such doping methods could also help to eliminate decreases in on/off ratio with increasing V_{DS}, as mentioned previously and illustrated in Fig. 2b and Supplementary Fig. 4a.

We find that standard models for silicon device technologies can capture macroscopic device behaviours. Figure 2b, c illustrates the level of agreement that can be achieved with a level-3 p-channel metal–oxide–semiconductor (PMOS) SPICE (simulation program for integrated circuits emphasis) model that uses a parallelly connected exponential current source controlled by both gate voltage and V_{DS} to mimic the electron tunnelling current. This level of compatibility with established simulation tools allows the use of existing sophisticated computer-aided design platforms developed for silicon integrated circuits.

As the first step towards large-scale integration, we modelled and then built 'universal' logic gates. Figure 3a shows a circuit diagram of a PMOS inverter with enhancement load. The inverter exhibits well-defined static voltage transfer characteristics, consistent with simulation, at a supply voltage of −5 V (Fig. 3b). The rise in output voltage with increasing positive input voltage is due to the ambipolar behaviour of the driving transistor. Maximum voltage gains of ~4, together with good noise immunity with a transition-region width of <0.8 V and a logic swing of >3 V are achieved, indicating that the inverter can be used to switch subsequent logic gates without losing logic integrity. Measuring their a.c. responses generated a

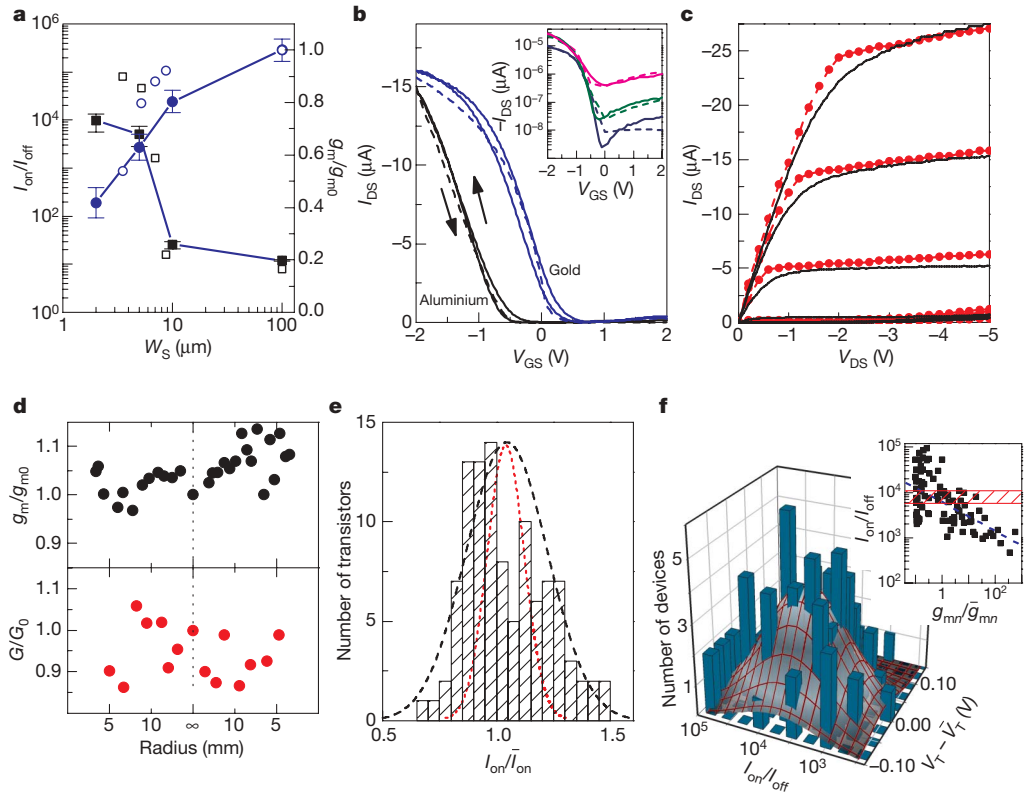


Figure 2 | Electrical properties of thin-film transistors that use SWNT network strips for the semiconductor, on thin plastic substrates. a, The measured (filled) and simulated (open) influence of the width of the strips (W_S) on the on/off ratio (I_{on}/I_{off}; black) and normalized transconductance (g_m/g_{m0}, where g_{m0} represents the response without strips; blue) of transistors with channel lengths of 100 μm. Error bars represent s.d. of $n = 6$ thin-film transistors. **b**, Measured (solid) and simulated (dashed) V_{GS}–I_{DS} characteristics of depletion-mode (blue) and enhancement-mode (black) SWNT thin-film transistors whose channel widths are 200 μm and whose channel lengths are 100 μm. $V_{DS} = -1$ V; I_{DS}, drain–source current; V_{GS}, gate–source voltage. Inset, V_{GS}–I_{DS} curve of the enhancement-mode device plotted on a logarithmic scale, with $V_{DS} = -0.5$ V (navy), −2 V (green), −5 V (magenta). **c**, Measured (black) and simulated (red) V_{DS}–I_{DS} characteristics of an enhancement-mode thin-film transistor (V_{GS} changed from −2 V to 2 V in steps of 0.5 V). **d**, Plots of g_m/g_{m0} for a thin-film transistor and G/G_0 (normalized voltage gain) for an inverter as functions of bend radius (g_{m0} and G_0 denote the responses in the unbent state). **e**, Histogram of I_{on} (measured at $V_{DS} = -0.2$ V; $\bar{I}_{on}$, averaged on-state current) with superimposed gaussian fitting for measured (dashed black) and simulated (dashed red) results. **f**, Two-dimensional histogram showing the correlation between the I_{on}/I_{off} (measured at $V_{DS} = -0.2$ V) and threshold voltage distributions ($\bar{V}_T$, averaged threshold voltage). Inset, correlation between I_{on}/I_{off} and normalized n-branch transconductance ($\bar{g}_{mn}$, averaged n-branch transconductance). The dashed blue line depicts the result of a linear fit. The hatched red area shows the distribution of I_{on}/I_{off} predicted by percolation models that do not explicitly account for the influence of source–drain contacts.

Bode magnitude plot closely resembling the characteristics of low-pass amplifiers, with operation in the kilohertz range even for devices with long channels ($L_C \approx 100\,\mu m$) and significant channel-width-normalized overlap capacitance ($\sim 40\,fF\,\mu m^{-1}$; Fig. 3c). The ability to achieve switching speeds in the kilohertz range with device geometries that are compatible with techniques such as screen printing is important for the potential use of such SWNT networks in low-cost, printed electronics[21]. By adding another driving transistor to the inverter, either in parallel with the pull-down transistor to incorporate OR logic (Fig. 3d, e) or in series to incorporate AND logic (Fig. 3g, h), it is possible to construct NOR and NAND logic gates, respectively. The output characteristics and simulation results are presented in Fig. 3f, i. Voltage amplification is observed in all cases.

All of these experimental and computational components can be used together to yield SWNT-based digital circuits (Fig. 4a). The largest circuit in this chip is a four-bit row decoder (Fig. 4b), designed using modelling tools and measured characteristics of stand-alone logic gates. This circuit incorporates 88 transistors, in four inverters and a NOR array, with the output of the inverter serving as one of the inputs for the NOR gate. The circuit diagram (Fig. 4c) is configured such that any given set of inputs only give one logic-'1' output. The input–output characteristics of the decoder are shown in Fig. 4d and Supplementary Fig. 6, which demonstrates its ability to decode a binary-encoded input of four data bits into sixteen individual data output lines, at frequencies in the kilohertz range. These results suggest that SWNT networks can form the basis for a potentially interesting and scalable alternative to conventional organic or other classes of semiconductors for flexible integrated circuitry applications. The development of optimized materials and solution-printing techniques for fabricating SWNT-based integrated circuits that achieve the performance levels reported here, together with further exploration of circuit- and systems-level implementation, represent some directions for future work.

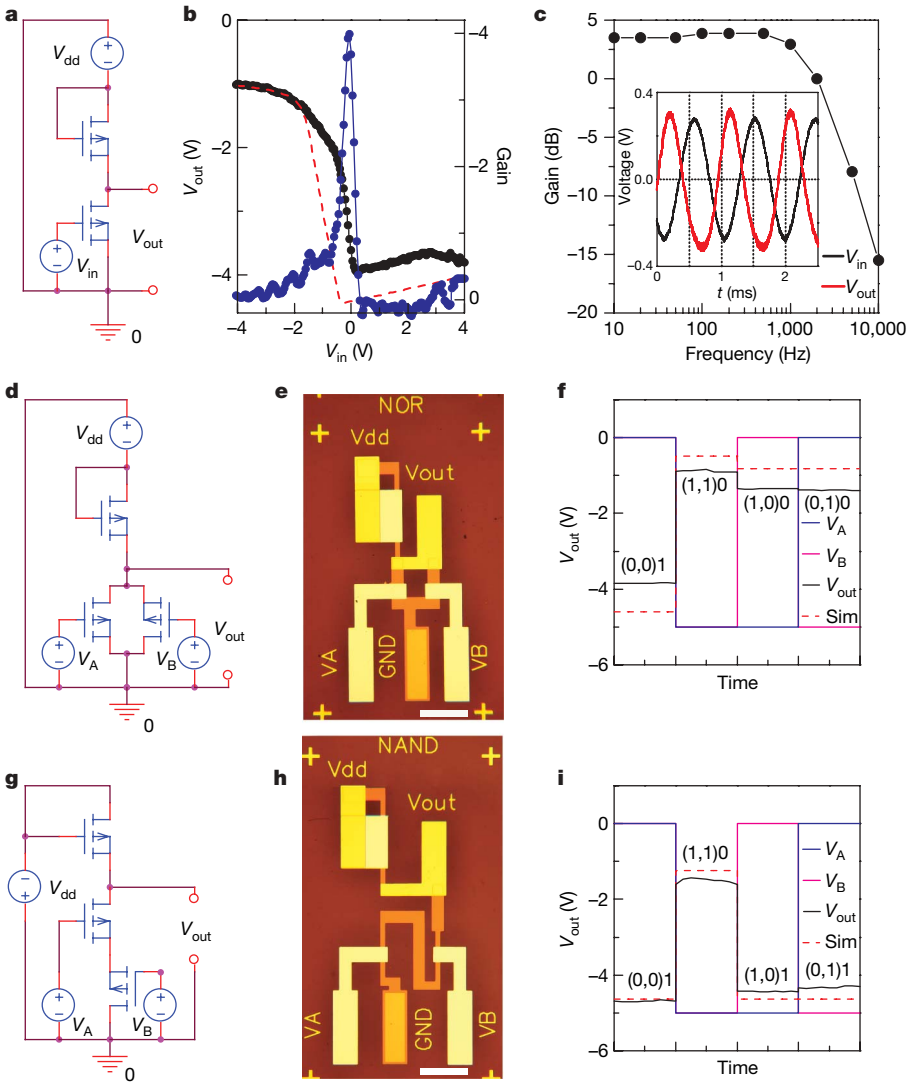


Figure 3 | Circuit diagram, optical micrographs, output–input characteristics and circuit simulation results for different logic gates. **a–c**, Inverter. **d–f**, NOR gate. **g–i**, NAND gate. We adopt a negative logic system. The V_{dd} applied to these logic gates is −5 V relative to GND. The logic-'0' and -'1' input signals of two terminals ($V_A \equiv$ VA and $V_B \equiv$ VB) of the NOR and NAND gates are driven by 0 V and −5 V, respectively. The logic-'0' and -'1' outputs of the NOR gate are −(0.88–1.39) V and –3.85 V, respectively. The logic-'0' and -'1' outputs of the NAND gate are –1.47 V and −(4.31–4.68) V, respectively. In **b**: black, V_{out}; blue, gain. In **f** and **i**, any specific combination of input–output signals is indicated as (logic address level inputs)logic address level ouput, and the timescales on the x axes are omitted because data collection involved the switching of voltage settings by hand. In **b**, **f** and **i**, dashed red lines represent circuit simulation results. Scale bars in **e** and **h**, 100 μm.

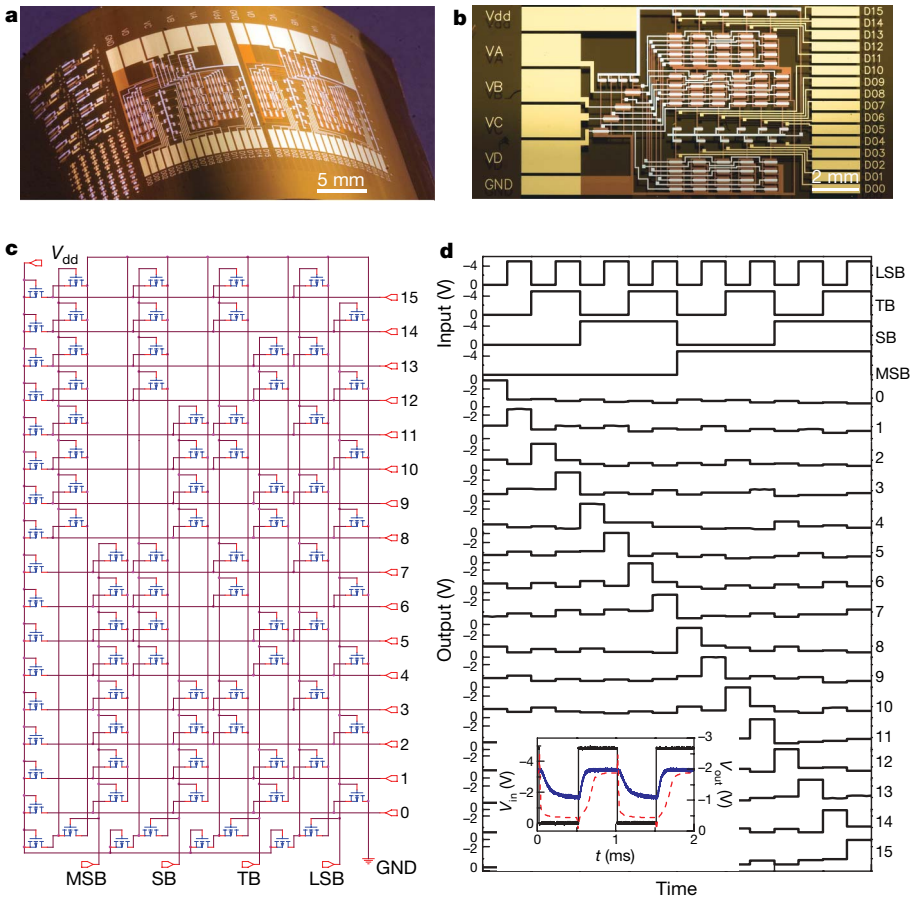


Figure 4 | Medium-scale integrated circuits based on SWNT network strips, on a thin plastic substrate. a, Optical image of a flexible SWNT integrated circuit chip bonded to a curved surface. **b**, Optical micrograph and **c**, circuit diagram of a four-bit row decoder with sixteen outputs (0–15). The bits are designated as most significant bit (MSB), second bit (SB), third bit (TB) and least significant bit (LSB). The V_{dd} applied was −5 V relative to GND. **d**, Characteristics of the four-bit decoder. In descending order, the first four traces are inputs, labelled LSB, TB, SB and MSB on the right-hand side; the remaining traces, labelled '0' to '15', show the output voltages of the sixteen outputs. Inset, measured (blue) and SPICE-simulated (red) dynamic response of one output line under a square-wave input pulse (black) at a clock frequency of 1 kHz.

METHODS SUMMARY

The process flow for fabricating SWNT integrated circuits on plastics is depicted in Supplementary Fig. 8. SWNTs were synthesized by chemical vapour deposition on silicon dioxide–silicon wafers and then etched into strips using an experimentally simple, optical soft lithography technique. Standard photolithography, electron-beam evaporation, gold wet chemical etching and oxygen plasma etching were used to pattern S–D electrodes and isolate each device. We then used a film of polyamic acid to encapsulate predefined S–D electrodes and SWNT networks on the growth wafers for transfer to a polyimide substrate coated with liquid polyurethane. Subsequent curing of the liquid polyurethane and polyamic acid completed the transfer process. Metal gates were defined on top of a high-capacitance dielectric (~40-nm) layer of hafnium dioxide. Vias and windows for probing were opened by wet etching (dipped into concentrated hydrofluoric acid aqueous solution) through patterned photoresist. Last, another level of interconnect metallization formed local interconnections defined previously with the gate and source–drain metal layers. All electrical measurements were carried out in air using a semiconductor parameter analyser (Agilent, 4155C). Alternating-current input was provided by a function generator (GW Instek, GFG-8219A) and output was read using a standard oscilloscope (Tektronix, TDS 3012B). The stick percolation simulations involved finite-size, first-principles two-dimensional numerical models based on generalized heterogeneous random network theory. Device and circuit simulation used the commercial software package HSPICE (Synopsis).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 29 January; accepted 20 May 2008.

1. Reuss, R. H. *et al.* Macroelectronics: Perspectives on technology and applications. *Proc. IEEE* **93,** 1239–1256 (2005).
2. Forrest, S. R. The path to ubiquitous and low-cost organic electronic appliances on plastic. *Nature* **428,** 911–918 (2004).
3. Gelinck, G. H. *et al.* Flexible active-matrix displays and shift registers based on solution-processed organic transistors. *Nature Mater.* **3,** 106–110 (2004).
4. Rogers, J. A. *et al.* Paper-like electronic displays: Large-area rubber-stamped plastic sheets of electronics and microencapsulated electrophoretic inks. *Proc. Natl Acad. Sci. USA* **98,** 4835–4840 (2001).
5. Someya, T. *et al.* Conformable, flexible, large-area networks of pressure and thermal sensors with organic transistor active matrixes. *Proc. Natl Acad. Sci. USA* **102,** 12321–12325 (2005).
6. Sekitani, T. *et al.* A large-area wireless power-transmission sheet using printed organic transistors and plastic MEMS switches. *Nature Mater.* **6,** 413–417 (2007).
7. Crone, B. *et al.* Large-scale complementary integrated circuits based on organic transistors. *Nature* **403,** 521–523 (2000).
8. Singh, T. B. & Sariciftci, N. S. Progress in plastic electronics devices. *Annu. Rev. Mater. Res.* **36,** 199–230 (2006).
9. Briseno, A. L. *et al.* Patterning organic single-crystal transistor arrays. *Nature* **444,** 913–917 (2006).
10. Blanchet, G. B., Loo, Y. L., Rogers, J. A., Gao, F. & Fincher, C. R. Large area, high resolution, dry printing of conducting polymers for organic electronics. *Appl. Phys. Lett.* **82,** 463–465 (2003).
11. Sirringhaus, H. *et al.* High-resolution inkjet printing of all-polymer transistor circuits. *Science* **290,** 2123–2126 (2000).
12. Avouris, P., Chen, Z. H. & Perebeinos, V. Carbon-based electronics. *Nature Nanotechnol.* **2,** 605–615 (2007).
13. Bradley, K., Gabriel, J. C. P. & Gruner, G. Flexible nanotube electronics. *Nano Lett.* **3,** 1353–1355 (2003).
14. Zhou, Y. X. *et al.* p-channel, n-channel thin film transistors and p-n diodes based on single wall carbon nanotube networks. *Nano Lett.* **4,** 2031–2035 (2004).
15. Snow, E. S., Campbell, P. M., Ancona, M. G. & Novak, J. P. High-mobility carbon-nanotube thin-film transistors on a polymeric substrate. *Appl. Phys. Lett.* **86,** 033105 (2005).

16. Seidel, R. *et al.* High-current nanotube transistors. *Nano Lett.* **4,** 831–834 (2004).
17. Kang, S. J. *et al.* High-performance electronics using dense, perfectly aligned arrays of single-walled carbon nanotubes. *Nature Nanotechnol.* **2,** 230–236 (2007).
18. Chimot, N. *et al.* Gigahertz frequency flexible carbon nanotube transistors. *Appl. Phys. Lett.* **91,** 153111 (2007).
19. Beecher, P. *et al.* Ink-jet printing of carbon nanotube thin film transistors. *J. Appl. Phys.* **102,** 043710 (2007).
20. Kocabas, C. *et al.* Experimental and theoretical studies of transport through large scale, partially aligned arrays of single-walled carbon nanotubes in thin film type transistors. *Nano Lett.* **7,** 1195–1202 (2007).
21. Chason, M., Brazis, P. W., Zhang, H., Kalyanasundaram, K. & Gamota, D. R. Printed organic semiconducting devices. *Proc. IEEE* **93,** 1348–1356 (2005).
22. Klauk, H., Zschieschang, U., Pflaum, J. & Halik, M. Ultralow-power organic complementary circuits. *Nature* **445,** 745–748 (2007).
23. Yoon, M. H., Yan, H., Facchetti, A. & Marks, T. J. Low-voltage organic field-effect transistors and inverters enabled by ultrathin cross-linked polymers as gate dielectrics. *J. Am. Chem. Soc.* **127,** 10388–10395 (2005).
24. Duan, X. F. *et al.* High-performance thin-film transistors using semiconductor nanowires and nanoribbons. *Nature* **425,** 274–278 (2003).
25. Kim, D. H. *et al.* Complementary logic gates and ring oscillators on plastic substrates by use of printed ribbons of single-crystalline silicon. *IEEE Trans. Electron Devices* **29,** 73–76 (2008).
26. Arnold, M. S., Green, A. A., Hulvat, J. F., Stupp, S. I. & Hersam, M. C. Sorting carbon nanotubes by electronic structure using density differentiation. *Nature Nanotechnol.* **1,** 60–65 (2006).
27. Chen, Z. H. *et al.* An integrated logic circuit assembled on a single carbon nanotube. *Science* **311,** 1735 (2006).
28. Shim, M., Ozel, T., Gaur, A. & Wang, C. J. Insights on charge transfer doping and intrinsic phonon line shape of carbon nanotubes by simple polymer adsorption. *J. Am. Chem. Soc.* **128,** 7522–7530 (2006).
29. Javey, A., Guo, J., Wang, Q., Lundstrom, M. & Dai, H. J. Ballistic carbon nanotube field-effect transistors. *Nature* **424,** 654–657 (2003).
30. Chen, J., Klinke, C., Afzali, A. & Avouris, P. Self-aligned carbon nanotube transistors with charge transfer doping. *Appl. Phys. Lett.* **86,** 123108 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank T. Banks, K. Colravy and D. Sievers for help with the processing. This material is based upon work supported by the US National Science Foundation (NIRT-0403489), the US Department of Energy (DE-FG02-07ER46471), Motorola, Inc., the Frederick-Seitz Materials Research Laboratory and the Center for Microanalysis of Materials (DE-FG02-07ER46453 and DE-FG02-07ER46471) at the University of Illinois. Q.C. acknowledges fellowship support from the Department of Chemistry at the University of Illinois. N.P., J.P.K., M.A. and K.R. acknowledge support from the Network for Computational Nanotechnology, which is supported by the National Science Foundation under cooperative agreement EEC-0634750. J.P.K. acknowledges fellowship support from the Intel Foundation.

Author Contributions Q.C., H.K. and J.A.R. designed the experiments. Q.C., H.K. and C.W. performed the experiments. Q.C., N.P., J.P.K., M.S., K.R., M.A.A. and J.A.R. analysed the data. Q.C. and J.A.R. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to J.A.R. (jrogers@uiuc.edu).

METHODS

Synthesis of SWNT networks. SWNT random networks were grown by chemical vapour deposition on silicon wafers with 100-nm-thick layers of thermal oxide. The process began with cleaning the SiO_2–Si wafer with piranha solution (a 3:1 volumetric mixture of concentrated sulphuric acid to 30% hydrogen peroxide solution). This process not only removed organic contaminants but also hydroxylated the re-oxidized SiO_2 surface, making it extremely hydrophilic to enable uniform deposition of catalyst[31]. This catalyst consisted of ferritin (Aldrich; diluted with de-ionized water at a volumetric ratio of 1:20 to control the density of catalyst) deposited onto the SiO_2–Si surface by adding methanol[32]. The wafer was then heated to 800 °C in a quartz tube to oxidize ferritin into iron oxide nanoparticles. After it had cooled down to room temperature, the quartz tube was flushed with a high flow of argon gas (1,500 s.c.c.m.) for cleaning and then heated up to 925 °C in hydrogen atmosphere (120 s.c.c.m.), which reduced iron oxide to iron. After the temperature had reached 925 °C, methane (1,500 s.c.c.m.) was released into the quartz tube as a carbon source while maintaining the hydrogen flow. Growth was terminated after 20 min, and the chamber was then cooled in hydrogen and argon flow. The density of the SWNT networks formed in this fashion was controlled by the dilution ratio of the ferritin solution, leaving the other aspects of the growth and processing unchanged.

Cutting strips into the SWNT networks with phase-shift lithography and reactive-ion etching. Elastomeric phase masks with depths of 1.8 μm, widths of 5 μm and periodicities of 10 μm were fabricated from relief structures defined by lithography and anisotropic etching through a casting and curing procedure[33]. AZ5214 photoresist, diluted with AZ1500 thinner in a 1:1 volumetric ratio, was spin-cast onto the SiO_2–Si wafer with SWNT networks at 5,000 r.p.m. and then baked at 95 °C for 1 min to afford a flat and solid 300-nm-thick photoresist layer. After cleaning the surface of phase mask with Scotch Tape, we placed it into conformal contact with the photoresist layer, flood exposed the resist by shining the i-line (365-nm) output of a mercury ultraviolet lamp through the mask, and then removed the mask. The SiO_2–Si substrate was then baked at 112 °C for another minute, followed by a flood exposure of ultraviolet light. Development in AZ MIF327 developer for 40 s created a regular array of submicrometre-wide spacings in the photoresist layer, with 5-μm periodicity. Photoresist strips of 5-μm width could also be generated by conventional photolithography with much wider spacings (~5 μm). Although large spacings lead to a reduction in effective channel width and an increase in parasitic capacitance, we used this technique instead of phase-shift lithography in fabricating the transistors used in the decoder circuits because conventional photolithography is easier to perform. We next used oxygen reactive-ion etching (200 mtorr, 20 s.c.c.m., O_2 flow, 100-W radio frequency power) to remove the exposed SWNTs. Last, the photoresist layer was removed by soaking in acetone for 1 h. Successfully using optical soft lithography to pattern the only sub-10-μm features in our circuits suggests the potential to use low-cost, low-resolution printing-like processes to define all features in the circuits[34].

S–D patterning and device isolation. A gold film (30 nm) was deposited by electron-beam evaporation (Temescal BJD 1800; base pressure of 3×10^{-6} torr) onto a SiO_2-Si substrate with predefined nanotube strips. We then used standard ultraviolet photolithography to pattern the S–D electrodes and interconnects using an etch-back scheme with a commercial wet etchant (Transene, TFA) to remove gold in exposed areas. After that we used oxygen reactive-ion etching (200 mtorr, 20 s.c.c.m., O_2 flow, 100-W radio frequency power) to remove SWNTs outside channel regions that were protected by a patterned layer of photoresist (Shipley 1805).

This step can also be carried out on the plastic substrate after transfer, which avoids the dimensional instability associated with polymer shrinkage during the curing process and device failure due to incomplete transfer of the S–D electrodes. However, it will lead to inferior device performance, owing to the synergetic effect of a smaller contact area between the S–D electrodes and the partially embedded SWNTs, as well as a smaller effective channel width when we use photolithography to define SWNT strips as described above on polymer surfaces that are relatively rough (in comparison with the surface roughness of the silicon wafers; Supplementary Fig. 3). Therefore, this approach is only used in fabricating row-decoder circuits, which process has the highest requirements on device yield.

Transfer-printing process. The transfer-printing process involved spin-casting (1,500 r.p.m., 60 s) polyamic acid (PAA, Aldrich) onto the SiO_2–Si wafer with SWNTs and S–D patterns, and then heating at 110 °C for 3 min to remove the solvent. On the target polyimide (PI) substrate (DuPont, Kapton E; thickness ~50 μm), we spin-cast (5,000 r.p.m., 60 s) a film of polyurethane (PU, NEA 121). Before this step, we thermally cycled the PI between 30 °C and 270 °C to improve its dimensional stability[35]. We laminated this PU-coated substrate on top of the PAA–SiO_2–Si wafer with the PU facing towards the PAA film, and applied pressure on the back of the wafer to remove air bubbles. Heating them together to 135 °C for 30 min thermally cured the PU film, thereby binding the PI substrate to the PAA film. Peeling off the PI substrate lifted the film of PU–PAA with embedded SWNT networks and S–D electrodes off the SiO_2–Si wafer, with one side of the S–D electrodes exposed. In the final step a vacuum oven (base pressure of 300 mtorr) with nitrogen flow (500 s.c.c.m.) was used to thermally cure the PAA to the PI through imidization reaction[36].

Gate dielectric deposition. The gate dielectric was deposited on top of the PAA after the latter had been cured to the PI. In the first step, 30 nm of HfO_2 was deposited by electron-beam evaporation (Temescal BJD 1800; base pressure of 2×10^{-6} torr) at a relatively low deposition rate (<0.5 Å s^{-1}) as measured by a quartz crystal thickness monitor. This layer served as a protective layer for SWNTs against highly reactive precursors used in a subsequent atomic layer-deposition (ALD) process[37]. After evaporation, the sample was transferred immediately to the ALD chamber to preserve the hydrophilicity of the freshly deposited HfO_2, which facilitates the growth of high-quality, pin-hole-free ALD film. The ALD HfO_2 film (12 nm) was deposited using a commercial ALD reactor (Cambridge Nanotech, Savannah 100). One ALD reaction cycle consists of one dose of water followed by a 5-s exposure and a 300-s purge, and then one dose of $Hf(NMe_2)_4$ followed by another 5-s exposure and a 270-s purge. During deposition, the nitrogen flow was fixed at 20 s.c.c.m. and the chamber temperature was set at 120 °C. The low deposition temperature prevents cracking of HfO_2 due to the mismatch of thermal expansion coefficients but requires very long purging time to remove excess precursors absorbed on the surface, to prevent chemical-vapour-deposition-type reactions in the chamber[38].

Via opening and gate/interconnect pattering. After the dielectric had been deposited, the gate pattern was defined in another photolithography step. A lift-off scheme was used to allow alignment of gate electrodes to the S–D electrodes using previously patterned alignment markers. Metal for the gate electrodes (120 nm aluminium or 2 nm chromium–120 nm gold) was deposited by electron-beam evaporation (Temescal BJD 1800; base pressure of 3×10^{-6} torr). In this metallization step (as well as the next step, for defining interlayer interconnects) two angled evaporations (incidence angle, 60°) with substrates placed at opposite orientations and a blanket evaporation (incidence angle, 90°) were performed to ensure that the metal layers covered the underlying surface topography, thereby avoiding open points that would otherwise form in the interconnect lines. In all cases, the deposition rate must be within 4–7 Å s^{-1}. If the evaporation rate is lower than 4 Å s^{-1}, accumulated heat can lead to cracking of the PU layer; if the evaporation rate is higher than 7 Å s^{-1}, the strain accumulated in the metal film can lead to defect formation in the lift-off process.

Following deposition, the lift-off was accomplished by soaking in acetone for 10 min, followed by a short ultrasonic treatment (30 s) to ensure that the lift-off process was complete. Because the SWNTs were covered by HFO_2, the ultrasonic treatment did not damage the nanotube network. (Prolonged acetone soaking can dissolve, at a low rate, the PI cured from PAA, owing, presumably, to incomplete imidization.) Contact pads for probing and vias for interlayer interconnects were exposed by photolithography using AZ 5214 photoresist. A hard bake (120 °C, 2 min) of the photoresist was performed before hydrofluoric acid etching (4 s in concentrated HF solution) of HfO_2 (ref. 39) to improve the adhesion between the photoresist and the HfO_2. We note here that in this step the gold pads patterned in the S–D layer under vias must be larger in size than the via holes to protect the PU from being etched by the hydrofluoric acid through acidolysis reaction. The interlayer interconnect (5 nm chromium–100 nm gold) was patterned using a lift-off process and photolithography. The patterning of gate electrodes and interconnects were carried out separately because (1) the predefined gate layer can also serve to protect the gate dielectric against possible defects existing in the photoresist mask layer, preventing the creation of pin holes in the channel region in the wet etching step, and (2) aluminium tends to form a poor contact with the gold S–D electrodes, possibly because of intermetallic formation[40], such that a different interconnect metal, such as the chromium–gold combination, was necessary when using aluminium gates. Finally, the completed device/circuit was aged in air for 24 h, and then thermally annealed at 120 °C for 30 min, to achieve stable operation.

Device and circuit characterizations. Direct-current measurements of SWNT transistors and circuits were carried out in air using a semiconductor parameter analyser (Agilent, 4155C), operated by Agilent Metrics I/CV Lite software (version 2.1) and GBIP communication. Triaxial and coaxial shielding was incorporated into a Signatone probe station to achieve a better signal-to-noise ratio. A precision LCR meter (Agilent, 4282A) was used for capacitance and impedance measurements. Alternating-current input signals were generated by a function generator (GW Instek, GFG-8219A). The output signals were measured using a standard oscilloscope (Tektronix, TDS 3012B).

Stick percolation simulation. We constructed a sophisticated first-principles numerical stick percolation model for the above random SWNT network by generalizing the random network theory[20,41,42]. The model randomly populates a two-dimensional grid with sticks of fixed length (L_S) and random orientation (θ) and determines I_{on} through the network by solving the percolating electron transport through individual sticks. In contrast to classical percolation, the SWNT network is a heterogeneous network: one-third of the carbon nanotubes are metallic and the remaining two-thirds are semiconducting. Because L_C and L_S here are much larger than the phonon mean free path, linear-response transport obviates the need to solve the Poisson equation explicitly. The system is well described by drift–diffusion theory within individual stick segments of this random stick network. The low-bias drift–diffusion equation, $J = q\mu n \mathrm{d}\varphi/\mathrm{d}s$ (where J is current density, q is carrier charge, μ is mobility, n is carrier density, φ is electropotential and s is length along the tube), when combined with the current continuity equation, $\mathrm{d}J/\mathrm{d}s = 0$, gives the non-dimensional potential φ_i along tube i as:

$$\frac{\mathrm{d}^2\varphi}{\mathrm{d}s^2} - C_{ij}(\varphi_i - \varphi_j) = 0$$

Here $C_{ij} = G_0/G_1$ is the dimensionless charge-transfer coefficient between tubes i and j at their intersection point. $G_0 \approx 0.1\, e^2/h$ and $G_1 = qn\mu/\Delta x$ are the mutual- and self-conductances of the tubes, respectively, and e is the elementary charge, h is Planck's constant and Δx is the grid spacing. The density of the random stick network is measured in area normalized by L_S, and the density of our SWNT network was ~40 according to scanning electron microscope measurements. The finite-length strips were simulated by imposing a reflecting boundary condition at the edge of each strip.

SPICE simulation. We described the behaviour of the SWNT thin-film transistors as that of a PMOS field-effect transistor parallelly connected with an exponential current source dependent on voltage (V_{GS} and V_{DS}). The PMOS field-effect transistor was modelled using a standard square-law model with channel-length modulation and S–D resistance effects. The exponential current source was used to mimic the ambipolar current ($I_{ambipolar}$), which led to an exponential increase in I_{off} with increasing V_{DS}. We expressed the exponential term in the form of a Taylor series

$$I_{\mathrm{ambipolar}} = K_n(V_{GS} + V_{G0})\left(1 + V_x + \frac{V_x^2}{2} + \cdots\right)$$

where K_n and V_{G0} are fitting parameters and V_x is defined as $V_x = V_{\mathrm{Threshold}} + \alpha V_{GS} - \beta V_{DS}$, and the first three terms were incorporated into the SPICE model. All fitting parameters were extracted from measured I–V characteristics (summarized in Supplementary Table 1). The channel-length scaling behaviour of these SWNT random network transistors can only be captured by our percolation modelling. The results of such models (for example, off-state resistances) can be used as inputs to the SPICE models to capture the full range of behaviours.

The above model was then used in designing and simulating digital logic circuits[43]. In transient simulation, load capacitance was calculated automatically from measured overlap capacitance ($330\,\mathrm{nF\,cm^{-2}}$) and gate capacitance ($80\,\mathrm{nF\,cm^{-2}}$) per unit area as well as estimated contact resistance ($11\,\mathrm{k\Omega}$), by the HSPICE program. Although the measured voltage responses of fabricated circuits agreed well with the design specifications, the current load responses showed behaviour only qualitatively similar to the simulation results (Supplementary Fig. 9). This deviation may be due to the relatively large batch-to-batch variations in device performance, which influenced the current load more significantly than they did the voltage responses.

31. Plummer, J. D., Deal, M. D. & Griffin, P. B. *Silicon VLSI Technology: Fundamentals, Practice and Modeling* Ch. 4 (Prentice Hall, Upper Saddle River, New Jersey, 2002).
32. Li, Y. M. *et al.* Growth of single-walled carbon nanotubes from discrete catalytic nanoparticles of various sizes. *J. Phys. Chem. B* **105,** 11424–11431 (2001).
33. Maria, J., Malyarchuk, V., White, J. & Rogers, J. A. Experimental and computational studies of phase shift lithography with binary elastomeric masks. *J. Vac. Sci. Technol. B* **24,** 828–835 (2006).
34. Menard, E. *et al.* Micro- and nanopatterning techniques for organic electronic and optoelectronic systems. *Chem. Rev.* **107,** 1117–1160 (2007).
35. Zhou, L. S., Jung, S. Y., Brandon, E. & Jackson, T. N. Flexible substrate micro-crystalline silicon and gated amorphous silicon strain sensors. *IEEE Trans. Electron Devices* **53,** 380–385 (2006).
36. Brekner, M. J. & Feger, C. Curing studies of a polyimide precursor. 2. Polyamic acid. *J. Polym. Sci. Pol. Chem.* **25,** 2479–2491 (1987).
37. Javey, A. *et al.* High-kappa dielectrics for advanced carbon-nanotube transistors and logic gates. *Nature Mater.* **1,** 241–246 (2002).
38. Hausmann, D. M., Kim, E., Becker, J. & Gordon, R. G. Atomic layer deposition of hafnium and zirconium oxides using metal amide precursors. *Chem. Mater.* **14,** 4350–4358 (2002).
39. Fujii, S., Miyata, N., Migita, S., Horikawa, T. & Toriumi, A. Nanometer-scale crystallization of thin HfO2 films studied by HF-chemical etching. *Appl. Phys. Lett.* **86,** 212907 (2005).
40. Philofsk, E. Intermetallic formation in gold-aluminum systems. *Solid State Electron.* **13,** 1391–1399 (1970).
41. Kumar, S., Murthy, J. Y. & Alam, M. A. Percolating conduction in finite nanotube networks. *Phys. Rev. Lett.* **95,** 066802 (2005).
42. Pimparkar, N. *et al.* Current-voltage characteristics of long-channel nanobundle thin-film transistors: A ''bottom-up'' perspective. *IEEE Electron Device Lett.* **28,** 157–160 (2007).
43. Rabaey, J. M. *Digital Integrated Circuits: A Design Perspective* (Prentice Hall, Upper Saddle River, New Jersey, 2002).

LETTERS

Archimedean-like tiling on decagonal quasicrystalline surfaces

Jules Mikhael[1], Johannes Roth[2], Laurent Helden[1] & Clemens Bechinger[1,3]

Monolayers on crystalline surfaces often form complex structures with physical and chemical properties that differ strongly from those of their bulk phases[1]. Such hetero-epitactic overlayers are currently used in nanotechnology and understanding their growth mechanism is important for the development of new materials and devices. In comparison with crystals, quasicrystalline surfaces exhibit much larger structural and chemical complexity leading, for example, to unusual frictional[2], catalytical[3] or optical properties[4,5]. Deposition of thin films on such substrates can lead to structures that may have typical quasicrystalline properties. Recent experiments have indeed showed 5-fold symmetries in the diffraction pattern of metallic layers adsorbed on quasicrystals[6,7]. Here we report a real-space investigation of the phase behaviour of a colloidal monolayer interacting with a quasicrystalline decagonal substrate created by interfering five laser beams. We find a pseudomorphic phase that shows both crystalline and quasicrystalline structural properties. It can be described by an archimedean-like tiling[8,9] consisting of alternating rows of square and triangular tiles. The calculated diffraction pattern of this phase is in agreement with recent observations of copper adsorbed on icosahedral $Al_{70}Pd_{21}Mn_9$ surfaces[10]. In addition to establishing a link between archimedean tilings and quasicrystals, our experiments allow us to investigate in real space how single-element monolayers can form commensurate structures on quasicrystalline surfaces.

Quasicrystals are unusual materials: they are aperiodic but retain true long-range order[11]. Although quasicrystalline structures have been theoretically also predicted in systems with a single type of particle[12,13], experimentally their spontaneous formation has been observed only in binary, ternary or even more complex alloys[14]. Accordingly, their surfaces exhibit a high degree of structural and chemical complexity and show unexpected mechanical, electrical and optical properties[15]. To understand the origin of those characteristics it is useful to disentangle structural and chemical aspects; this can be achieved by growing single-element monolayers to quasicrystalline surfaces[16,17]. Apart from adding to our understanding of how quasicrystalline properties can be transferred to such monolayers[18], this approach might permit the fabrication of materials with previously unobserved properties. Heteroepictatic growth experiments on decagonal and icosahedral surfaces did indeed show the formation of Bi and Sb monolayers with a high degree of quasicrystalline order as determined by low-energy electron diffraction and elastic helium-atom scattering experiments[6,18]. In comparison with reciprocal space studies, it was only recently that scanning tunnelling microscopy permitted an atomic resolution of the adsorbate morphology[7]. Even then, however, it was difficult to relate the structure of the adsorbate to that of the underlying substrate.

Here we report an experimental study of the phase behaviour of a two-dimensional colloidal monolayer in the presence of a quasicrystalline substrate potential. The quasicrystalline substrate potential is created by the interference of five laser beams in the sample plane (Fig. 1a, b). As a result of optical gradient forces[19], this light pattern acts as a decagonal quasicrystalline surface for the colloidal particles[20,21]. Figure 1b shows the light intensity distribution in the sample plane, which displays maxima arranged in pentagons of different

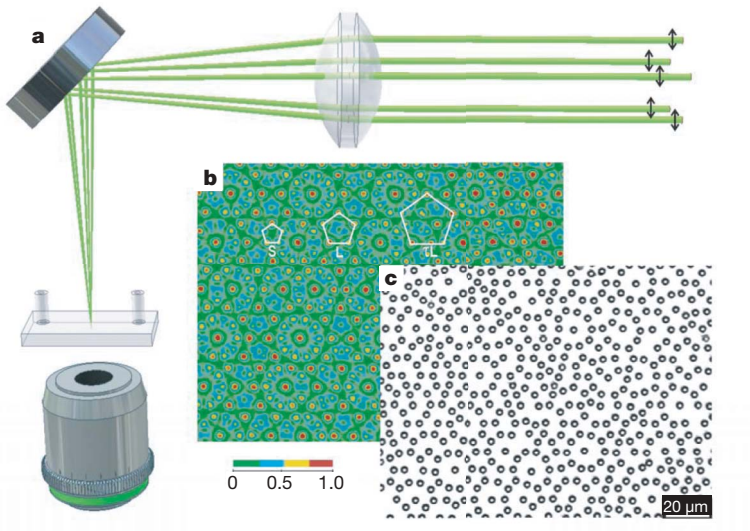


Figure 1 | Experimental realization of colloidal quasicrystals. a, Five linearly polarized (polarization as indicated by arrows) parallel laser beams forming a regular pentagon are focused with an achromatic lens into a thin sample cell. **b**, Experimentally determined intensity distribution of the interference pattern, which acts as a substrate potential for the colloids. The pattern displays a decagonal symmetry and the predominating motifs are pentagons (indicated in white) with sides of different lengths related by the golden ratio $\tau = L/S$. Here $S = 5.64\,\mu m$ and $L = 9.13\,\mu m$. The colour coding of the intensity field ranges from green to red and reflects the variation in potential well depth. **c**, Configuration of colloidal particles at a density of $\Phi = 0.0264\,\mu m^{-2}$ exposed to a decagonal substrate interference pattern.

[1]2. Physikalisches Institut, Universität Stuttgart, Pfaffenwaldring 57, 70569 Stuttgart, Germany. [2]Institut für Theoretische und Angewandte Physik, Universität Stuttgart, Pfaffenwaldring 57, 70569 Stuttgart, Germany. [3]Max-Planck-Institut für Metallforschung, Heisenbergstrasse 3, 70569 Stuttgart, Germany.

sizes whose side lengths and heights are related by the golden ratio $\tau = (1+\sqrt{5})/2 = 1.618\ldots$. The substrate strength can be continuously adjusted by the laser intensity I_0 (ref. 22). As a colloidal suspension we used highly charged polystyrene spheres of radius $R = 1.45\,\mu m$ suspended in water in a silica cuvette with a height of 200 μm. Particles at distance r interact through a repulsive screened electrostatic-pair potential $u(r) \propto \frac{1}{r}\exp(-\kappa r)$ whose range depends on the Debye screening length κ^{-1} given by the ion concentration of the suspension[23]. Particle positions are observed in real space with digital video microscopy, which allows us to determine their positions relative to the substrate potential with a precision of about 50 nm (Fig. 1c). From the particle coordinates we obtain the particle density distribution $\rho(x,y)$ and the two-dimensional structure factor $S(q_x,q_y)$, the latter being equivalent to the diffraction pattern. In addition we calculated the projection of the structure factor on the q_x axis, $\bar{S}(q_x)$, which corresponds to a slice in the real-space structure along the projection direction[14].

Figure 2a–d shows how $\rho(x,y)$ and $S(q_x,q_y)$ of a colloidal monolayer with a particle density $\Phi = 0.0465\,\mu m^{-2}$ and $\kappa^{-1} \approx 160$ nm changes when the substrate strength of the decagonal substrate is increased. In absence of a substrate potential ($I_0 = 0$) the system crystallizes as shown in Fig. 2a, b. The observed structure can be described by a triangular lattice (green tiles) whose vertices are defined by the maxima of $\rho(x,y)$. Apart from some defects, each vertex is surrounded by six triangular tiles forming a (3^6)-vertex type. The crystalline structure is confirmed by the diffraction pattern, which has 6-fold coordinated spots leading to periodically spaced peaks in the projected diffraction pattern $\bar{S}(q_x)$ (Fig. 2b). The intensity of the peaks decreases with increasing diffraction order, reflecting the thermal motion of particles[24]. In the presence of a quasicrystalline light field, the particles also interact with the corresponding surface potential, and the equilibrium structure will change. To establish equilibrium conditions, the laser intensity I_0 was increased at a rate much smaller than the typical relaxation time of the colloidal system. At low I_0, the electrostatic colloidal repulsion dominates over the colloid–substrate interaction and the crystalline structure remains mainly intact. In agreement with numerical simulations of weakly adsorbed atomic systems, we observe the alignment of crystalline domains along the 5-fold directions of the quasicrystalline substrate[25].

For $I_0 > 1.3\,\mu W\,\mu m^{-2}$, however, we observe a structure that shows neither a triangular nor a decagonal symmetry. As an example we show the situation for $I_0 = 2\,\mu W\,\mu m^{-2}$ (Fig. 2c, d). In contrast with a triangular structure, in which the nearest-neighbour distance distribution has a single peak, here it is bimodal (data not shown). The structure is well characterized by a tiling composed of squares (blue) and triangles (green), with the vertex–vertex bonds being selected by their length and angle (for details see Methods). The tiling structure has rows of triangles and squares mainly aligned in one direction with some intrusions at an angle of 72°. The direction of the rows varied between different experiments but always corresponded to one of the five orientations given by the substrate potential. The peaks in $S(q_x,q_y)$ are periodically spaced along the q_x direction (as clearly seen in the projection $\bar{S}(q_x)$), whereas in the q_y direction their distance is close to the golden ratio τ (see vertical spacing of white arrows). This is a clear signature of quasicrystalline order along the y direction. Obviously, the competition between the colloid–colloid repulsion and their interaction with the quasicrystalline substrate leads to a phase that has both periodic and quasicrystalline structural properties.

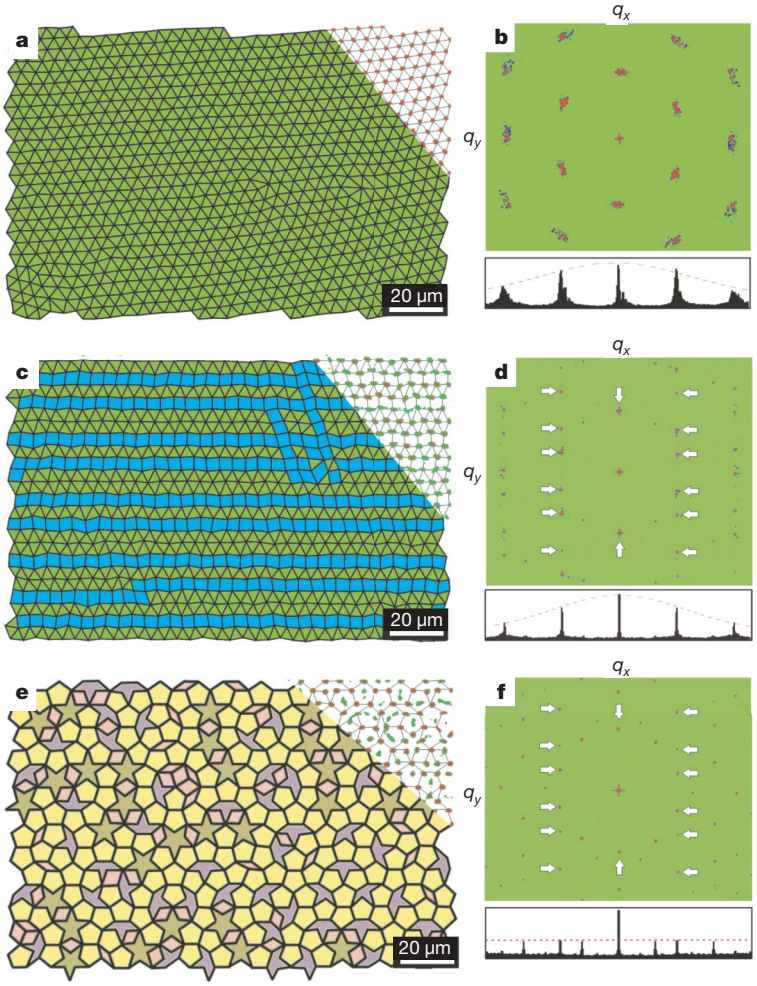


Figure 2 | Real and reciprocal space structure of the adsorbate. **a**, **b**, Crystalline phase ($I_0 = 0$, $\kappa^{-1} \approx 160$ nm). **a**, Particle density distribution (shown in the upper right corner) and tiling with triangles. Increasing particle density is labelled from white through green to red. **b**, Diffraction pattern and projection $\bar{S}(q_x)$. **c**, **d**, Intermediate phase ($I_0 = 2\,\mu W\,\mu m^{-2}$, $\kappa^{-1} \approx 160$ nm). **c**, Particle density and tiling of colloidal monolayer subjected to a decagonal light lattice. The tiling consists of rows containing triangles (green) and squares (blue). **d**, $S(q_x,q_y)$ and $\bar{S}(q_x)$. The arrows indicate diffraction peaks also found in the quasicrystalline phase in **f**. **e**, **f**, Quasicrystalline phase ($I_0 = 2\,\mu W\,\mu m^{-2}$, $\kappa^{-1} \approx 10$ nm). **e**, Particle density and tiling composed of pentagons (yellow), rhombuses (pink), crowns (violet) and pentagonal stars (brown). **f**, $S(q_x,q_y)$ with 10-fold symmetry and $\bar{S}(q_x)$ with peak spacing given by τ.

Before analysing this new intermediate phase in more detail, we discuss the observations when the colloid–substrate interaction dominates. This is achieved by increasing the ionic strength of the suspension, which greatly reduces the repulsion between the colloids. For $I_0 = 2\ \mu W\ \mu m^{-2}$ and Debye screening lengths $\kappa^{-1} \leq 95 \pm 15$ nm, the particle density distribution changes greatly in comparison with Fig. 2c, and the maxima ($\rho \geq 0.7\rho_{max}$) now follow a tiling consisting of prototiles shaped like rhombuses, pentagons, crowns and stars (Fig. 2e). Such tiles are known from the P1 Penrose tiling and have also been applied successfully to the 5-fold surface of AlPdMn (ref. 26). The corresponding diffraction pattern shows perfect quasicrystalline order and thus mimics the geometry of the underlying light potential (Fig. 2f). In the q_y direction the diffraction spots (arrows) coincide with those of the intermediate phase but, unlike in Fig. 2d, here $S(q_x,q_y)$ has a 10-fold rotational symmetry. As expected, $\bar{S}(q_x)$ consists of peaks whose distances are not equal but are given by τ. In contrast to Fig. 2d, the intensity of the peaks of $\bar{S}(q_x)$ is constant. This is consistent with the fact, that as a result of the strong substrate interactions the thermal motion of the particles becomes reduced[24].

The structure of the intermediate phase is remarkably similar to one of the 11 archimedean tilings[8] first introduced by Kepler in 1619. Currently, there is renewed interest in archimedean tilings as candidates for photonic crystals[9]. In contrast to the five two-dimensional Bravais lattices each described by identical tiles (being the corresponding unit cell), archimedean tilings may be composed of more than one, but regular, tile. Those tiles are arranged in such a way that only one vertex type exists. Figure 3a shows an example of an archimedean tiling consisting of alternating rows of triangular and square tiles. Because each vertex is surrounded by three triangles and two squares, this leads to a $(3^3.4^2)$-vertex type. Although the structure is strictly periodic, it has marked similarities with quasicrystals. First, every vertex has five nearest neighbours at equal distances that form an irregular pentagon. Second, the structure of the archimedean tiling is equivalent to an oblique lattice (red lines) with a two-atomic basis. The oblique angle $\gamma = 75°$ is close to the value of 72° on decagonal substrates. Accordingly, when superimposing ideal pentagons (white lines) on the archimedean tiling, their vertices (and the centre of the bigger one) agree almost perfectly with the vertex positions. The height ratio of these pentagons equals the golden ratio τ.

To understand how the observed intermediate phase forms on the quasicrystalline substrate, we plotted the contours of the tiles taken from Fig. 2c (black lines) on top of the decagonal intensity distribution of the laser field (Fig. 3b). The deepest potential wells of the substrate coincide with vertices and thus show quasicrystalline order. For our particle density, this applies to about half of all vertex positions. The other vertices are located at sites with weak or vanishing substrate interactions, and their configuration is dominated by electrostatic particle repulsion. They therefore assemble in such a way that their nearest-neighbour distance is fairly uniform. As a result, vertices partly adopt a 5-fold rotational symmetry but simultaneously seek to achieve equal nearest-neighbour distances. Both aspects are ideally supported by the archimedean tiling.

Because archimedean tilings are strictly periodic, they can be only locally commensurate with quasicrystalline substrates; disruptions at larger length scales must occur. Indeed, the tiling of the intermediate phase in Fig. 2c shows additional interstitial rows of triangles (that is, double triangular rows). As a result, two vertex types, namely $(3^3.4^2)$ and (3^6), arise. Such structures are referred to as archimedean-like tilings[9]. One would expect that the spacing of those interstitial rows corresponds to a Fibonacci sequence, taking into account the long-range quasicrystalline order along one direction. This is consistent with the structure observed in Fig. 2c. The origin of the two characteristic length scales of such a Fibonacci chain is due to the height of small and big pentagons as shown in Fig. 3b. Whenever two pentagons adjoin, the periodic sequence of triangular and squared rows becomes disrupted and an additional row of triangular tiles is inserted.

We also investigated whether the intermediate phase is stable with other parameters. Figure 3c shows the result when a colloidal layer with density $\Phi = 0.0307\ \mu m^{-2}$ is exposed to a decagonal substrate potential whose characteristic length scales are decreased to about 70% compared with Fig. 3b. In contrast to the above, here the number of deep potential wells provided by the substrate is larger than the number of vertices. Nevertheless, the system's structure is again well described by an archimedean-like tiling and even though the magnitude of particle fluctuations relative to the substrate is stronger than in Fig. 3b, the diffraction pattern (Fig. 3d) agrees well with that in Fig. 2d. This is more clearly seen in the projection of $S(q_x,q_y)$, which shows again equidistant peaks with decreasing intensity and suggests that the intermediate phase forms for a wider range of parameters.

Because the phase behaviour of colloidal monolayers on surfaces is similar to that of atomic systems[27], we expect that structures comparable to those in Fig. 3 should occur in atomic adsorbates on quasicrystalline surfaces. Recent experiments with thin copper films

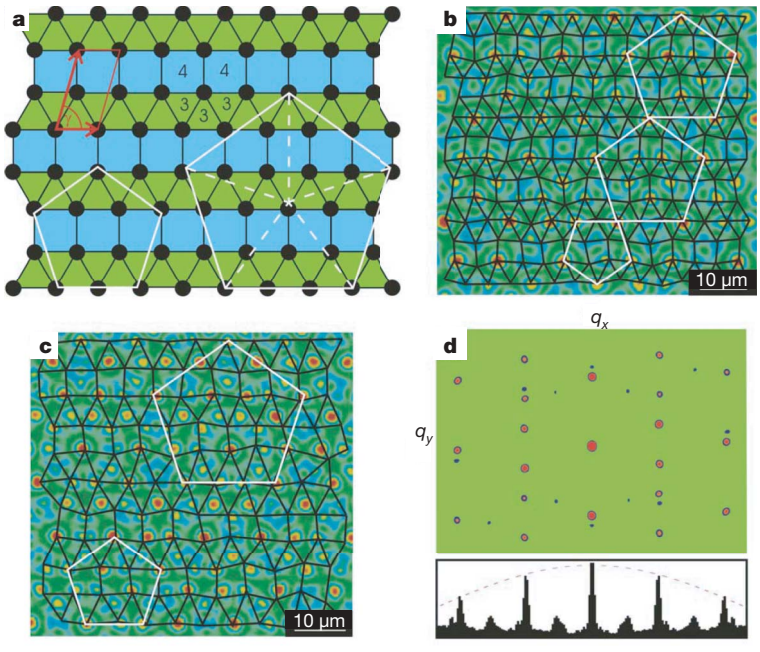


Figure 3 | Substrate–adsorbate correlations. **a**, Ideal archimedean tiling with $(3^3.4^2)$ vertex type. The vertices and centres of the two pentagons (white) fit the lattice sites almost perfectly. The structure can be also represented by an oblique lattice with two atoms per unit cell (red). **b**, Tiling of the intermediate phase superimposed on the laser intensity distribution of the decagonal interference pattern ($S = 5.65\ \mu m$, $L = 9.14\ \mu m$). Particles are partly located at deep minima in substrate potential, thus showing quasicrystalline order as indicated by the pentagons (white). Other colloids are located at interstitial sites with weak substrate interactions. **c**, As in **b** except for $\Phi = 0.0307\ \mu m^{-2}$ and $S = 4.10\ \mu m$, $L = 6.63\ \mu m$. **d**, Structure factor and $\bar{S}(q_x)$ of **c**, which is similar to that in Fig. 2d.

deposited on the 5-fold surface of icosahedral AlPdMn quasicrystals reveal that above a few monolayers the copper atoms are arranged in rows spaced in a Fibonacci sequence[10,28]. The atomic positions relative to the substrate have not yet been identified. However, although the atomic pair interactions in these experiments are more complex than in colloidal systems, the diffraction pattern of the copper film is almost identical to that of our intermediate phase (see Supplementary Fig. 1). This close resemblance suggests that the intermediate phase does not necessarily require complex substrate interactions but is driven by geometrical considerations and thus might also be observed for other adsorbate/quasicrystal combinations.

Our experiments show that colloidal systems on decagonal light patterns allow us to understand the equilibrium structure of monolayers on quasicrystalline surfaces. This approach can be also extended to investigate dynamical processes on quasicrystalline surfaces. By introducing phase shifts between the interfering laser beams, phason[29] or phonon modes can be induced in the substrate. These elementary excitations are important for the three-dimensional growth of quasicrystals and it will be interesting to study how such substrate excitations modify the behaviour of adsorbed thin films.

METHODS SUMMARY

The particle density Φ is adjusted by an optical fence created by an additional laser beam ($\lambda = 488$ nm), which was scanned around the central region of the sample to create a boundary box[30]. Its size can be continuously adjusted by a pair of computer-controlled scanned mirrors (Scanlab AG). With this technique, the particle density can be adjusted within a precision of about 1%. An additional laser beam ($\lambda = 514$ nm), which is vertically incident on the sample cell from above, pushes the colloids towards the negatively charged substrate and reduces vertical particle fluctuations to less than 5% of the particle radius. The system can therefore be considered to be two-dimensional.

As colloidal particles we used polystyrene particles with a radius R of 1.45 μm, a polydispersity of 4% and a negative surface charge density of 9.8 μC cm^{-2} (batch no. 1212; IDC). To reduce the ionic strength, we deionized the sample cell with a deionization circuit containing a vessel of ion-exchange resin, an electrical conductivity probe to measure the ionic concentration, and a peristaltic pump. Afterwards the suspension was inserted into the cell, which was then sealed. The Debye screening length was determined from the measured pair correlation function in the absence of the quasicrystalline lattice[30].

For the identification of phases we performed a Delaunay triangulation of the vertices as obtained by the maxima of $\rho(x,y)$ to identify next-neighbour bonds. The bond length distribution depends sensitively on the different phases, as follows: first, crystal, a monomodal distribution with a peak located at the mean particle distance; second, intermediate, a bimodal distribution with the ratio of the peak positions close to $\sqrt{2}$ and none of the peaks located at S and L (when the longer bonds are removed from the triangulation, one obtains square and triangular tiles as shown in Fig. 2c); and third, quasicrystalline, a bimodal distribution with peaks at S and L.

Received 29 November 2007; accepted 6 May 2008.

1. Barth, J. V., Costantini, G. & Kern, K. Engineering atomic and molecular nanostructures at surfaces. *Nature* **437,** 671–679 (2005).
2. Park, J. Y. *et al.* High frictional anisotropy of periodic and aperiodic directions on a quasicrystal surface. *Science* **309,** 1354–1356 (2005).
3. Tsai, A. P. & Yoshimura, M. Highly active quasicrystalline Al-Cu-Fe catalyst for steam reforming of methanol. *Appl. Catal. A* **214,** 237–241 (2001).
4. Zoorob, M. E., Charlton, M. D. B., Parker, G. J., Baumberg, J. J. & Netti, M. C. Complete photonic bandgaps in 12-fold symmetric quasicrystals. *Nature* **404,** 740–743 (2000).
5. Matsui, T., Agrawal, A., Nahata, A. & Vardeny, Z. V. Transmission resonances through aperiodic arrays of subwavelength apertures. *Nature* **446,** 517–521 (2007).
6. Franke, K. J. *et al.* Quasicrystalline epitaxial single element monolayers on icosahedral Al-Pd-Mn and decagonal Al-Ni-Co quasicrystal surfaces. *Phys. Rev. Lett.* **89,** 156104 (2002).
7. Sharma, H. R., Shimoda, M., Ross, A. R., Lograsso, T. A. & Tsai, A. P. Real-space observation of quasicrystalline Sn monolayer formed on the fivefold surface of icosahedral Al-Cu-Fe quasicrystal. *Phys. Rev. B* **72,** 045428 (2005).
8. Pearce, P. *Structure in Nature is a Strategy for Design* (MIT Press, Cambridge, MA, 1978).
9. David, S., Chelnokov, A. & Lourtioz, J. M. Isotropic photonic structures: Archimedean-like tilings and quasi-crystals. *IEEE J. Quantum Electron.* **37,** 1427–1434 (2001).
10. Ledieu, J. *et al.* Copper adsorption on the fivefold $Al_{70}Pd_{21}Mn_9$ quasicrystal surface. *Phys. Rev. B* **72,** 035420 (2005).
11. Shechtman, D., Blech, I., Gratias, D. & Cahn, J. W. Metallic phase with long-range orientational order and no translational symmetry. *Phys. Rev. Lett.* **53,** 1951–1953 (1984).
12. Engel, M. & Trebin, H.-R. Self-assembly of monoatomic complex crystals and quasicrystals with a double-well interaction potential. *Phys. Rev. Lett.* **98,** 225505 (2007).
13. Keys, A. S. & Glotzer, S. C. How do quasicrystals grow? *Phys. Rev. Lett.* **99,** 235503 (2007).
14. Janot, C. *Quasicrystals—A Primer* (Oxford Univ. Press, New York, 1994).
15. Dubois, J. M. Quasicrystals. *J. Phys. Condens. Matter* **13,** 7753–7762 (2001).
16. Fournee, V. *et al.* Nucleation and growth of Ag films on a quasicrystalline AlPdMn surface. *Phys. Rev. B* **67,** 033406 (2003).
17. Curtarolo, S., Setyawan, W., Ferralis, N., Diehl, R. D. & Cole, M. W. Evolution of topological order in Xe films on a quasicrystal surface. *Phys. Rev. Lett.* **95,** 136104 (2005).
18. Sharma, H. R., Shimoda, M. & Tsai, A. P. Quasicrystal surfaces: structure and growth of atomic overlayers. *Adv. Phys.* **56,** 403–464 (2007).
19. Ashkin, A. Optical trapping and manipulation of neutral particles using lasers. *Proc. Natl Acad. Sci. USA* **94,** 4853–4860 (1997).
20. Burns, M. M., Fournier, J. M. & Golovchenko, J. A. Optical matter—crystallization and binding in intense optical fields. *Science* **249,** 749–754 (1990).
21. Roichman, Y. & Grier, D. G. Holographic assembly of quasicrystalline photonic heterostructures. *Opt. Express* **13,** 5434–5439 (2005).
22. Bechinger, C., Brunner, M. & Leiderer, P. Phase behavior of two-dimensional colloidal systems in the presence of periodic light fields. *Phys. Rev. Lett.* **86,** 930–933 (2001).
23. Yethiraj, A. Tunable colloids: control of colloidal phase transitions with tunable interactions. *Soft Matter* **3,** 1099–1115 (2007).
24. Guinier, A. *X-ray Diffraction—In Crystals, Imperfect Crystals and Amorphous Bodies* (Dover, New York, 1994).
25. Bilki, B., Erbudak, M., Mungan, M. & Weisskopf, Y. Structure formation of a layer of adatoms on a quasicrystalline substrate: Molecular dynamics study. *Phys. Rev. B* **75,** 045437 (2007).
26. Ledieu, J. *et al.* Tiling of the fivefold surface of $Al_{70}Pd_{21}Mn_9$. *Surf. Sci.* **492,** L729–L734 (2001).
27. Yethiraj, A. Tunable colloids: control of colloidal phase transitions with tunable interactions. *Soft Matter* **3,** 1099–1115 (2007).
28. Ledieu, J. *et al.* Pseudomorphic growth of a single element quasiperiodic ultrathin film on a quasicrystal substrate. *Phys. Rev. Lett.* **92,** 135507 (2004).
29. Freedman, B., Lifshitz, R., Fleischer, J. W. & Segev, M. Phason dynamics in nonlinear photonic quasicrystals. *Nature Mater.* **6,** 776–781 (2007).
30. Brunner, M., Bechinger, C., Strepp, W., Lobaskin, V. & v Grünberg, H. H. Density-dependent pair-interactions in 2D colloidal suspensions. *Europhys. Lett.* **58,** 926–932 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. Baumgartl, S. Rausch, H.-H. v. Grünberg, M. Schmiedeberg and H. Stark for technical support and helpful discussions. This work is financially supported by the Deutsche Forschungsgemeinschaft.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to C.B. (c.bechinger@physik.uni-stuttgart.de).

LETTERS

Near-surface wetland sediments as a source of arsenic release to ground water in Asia

Matthew L. Polizzotto[1], Benjamin D. Kocar[1], Shawn G. Benner[2], Michael Sampson[3] & Scott Fendorf[1]

Tens of millions of people in south and southeast Asia routinely consume ground water that has unsafe arsenic levels[1,2]. Arsenic is naturally derived from eroded Himalayan sediments, and is believed to enter solution following reductive release from solid phases under anaerobic conditions. However, the processes governing aqueous concentrations and locations of arsenic release to pore water remain unresolved, limiting our ability to predict arsenic concentrations spatially (between wells) and temporally (future concentrations) and to assess the impact of human activities on the arsenic problem[3–9]. This uncertainty is partly attributed to a poor understanding of groundwater flow paths altered by extensive irrigation pumping in the Ganges-Brahmaputra delta[10], where most research has focused. Here, using hydrologic and (bio)geochemical measurements, we show that on the minimally disturbed Mekong delta of Cambodia, arsenic is released from near-surface, river-derived sediments and transported, on a centennial timescale, through the underlying aquifer back to the river. Owing to similarities in geologic deposition, aquifer source rock and regional hydrologic gradients[11–15], our results represent a model for understanding pre-disturbance conditions for other major deltas in Asia. Furthermore, the observation of strong hydrologic influence on arsenic behaviour indicates that release and transport of arsenic are sensitive to continuing and impending anthropogenic disturbances. In particular, groundwater pumping for irrigation, changes in agricultural practices, sediment excavation, levee construction and upstream dam installations will alter the hydraulic regime and/or arsenic source material and, by extension, influence groundwater arsenic concentrations and the future of this health problem.

There is general agreement that arsenic contamination in the ground water of south and southeast Asia is a consequence of arsenic release from sediment solids into pore water under anaerobic conditions and ensuing microbially mediated Fe(III) and As(V) reduction[3,7–9,11,14–18]. However, the location within the sediment profile, the time period, and the influence of hydrology on arsenic release remain unresolved; such information is crucial for defining remedial responses and predicting future arsenic concentrations.

We have formulated a coupled hydrologic and biogeochemical model of arsenic release and transport within the arsenic-contaminated Mekong River floodplain of Cambodia that pre-dates the complex influences of widespread irrigation. Our 50 km^2 field area includes >100 installed wells, lysimeters and surface water sites and is typical of the region, with native wetlands contained between delta river branches and a grey sand aquifer (≥40 m thick) overlain by a clay/silt layer (5–20 m thick); based on the wetland/river geometry, the groundwater system can be approximated by a two-dimensional cross-section perpendicular to the river. Although the Mekong delta system in Cambodia has similar depositional history, regional hydrology and biogeochemical conditions to other arsenic-contaminated deltaic aquifers of Asia[14] (Supplementary Information), land use alteration, inclusive of irrigation, is minimal. Thus, the hydrology of our system remains governed by natural rather than anthropogenic processes.

As with other Asian river deltas, regional hydrology is controlled by seasonal river fluctuations of ~8 m (Fig. 1). Groundwater levels mimic river levels, and the fluctuation amplitude decreases with distance from the Mekong River, indicating the strong influence of the river on the floodplain aquifer. The hydraulic gradient between the aquifer and river inverts annually: during the rising river stage, the subsurface gradient is from the river to the floodplain aquifer, but during the falling river stage, the gradient is towards the river (Supplementary Figs 2 and 3). Changes in surface water levels are clearly distinct from those observed at depth, producing temporally variable but strong vertical gradients between the surface water and underlying aquifer.

Despite seasonal hydraulic gradient inversions, a net annual head difference of 1.4 m exists between the wetlands and the river, producing a net downward gradient from the wetlands to the aquifer of 0.05–0.07 m m^{-1} and a net horizontal gradient from the aquifer to the river of 7×10^{-5} m m^{-1} (Supplementary Information); these findings agree with those predicted for Bangladesh before irrigation pumping[10,19]. The calculated downward flux through the confining clay layer is consistent with the independently calculated net horizontal flux to the river, revealing annual water balance between inflow and outflow. These observations indicate a groundwater travel time from the wetlands to the river in the range of 200–2,000 yr, and these results are supported by numerical modelling (Supplementary Figs 4 and 5). The age of the aquifer, and associated sedimentary organic carbon, is greater than 6,000 yr, based on both ^{14}C dating and regional geologic history[20–22], and, accordingly, the aquifer has been flushed by at least 3–30 pore volumes.

Arsenic concentrations within the aquifer range from 100 μg l^{-1} to >1,000 μg l^{-1}, and average ~500 μg l^{-1} (Supplementary Table 1). Groundwater flow having effectively flushed the aquifer, either an upstream source of arsenic must exist or arsenic must be continually released from aquifer solids—or a combination of both must occur—for arsenic to persist within the aquifer. Based on our yearly aquifer groundwater fluxes and average aqueous As concentration of 500 μg l^{-1}, $(2–20) \times 10^5$ kg of arsenic is removed from the aquifer system within our field area annually via transport to the river. ^{14}C dates indicate an average clay layer deposition rate of ~1–3.3 mm yr^{-1} over the past 6,000 yr, yielding a delivery rate of approximately $(6–20) \times 10^5$ kg of arsenic to the field area annually (Supplementary Information). Thus, quantities of arsenic influx (via sediment deposition) and efflux (aqueous transport from aquifer to river) are comparable, indicating that release of arsenic from solids

[1]School of Earth Sciences, Stanford University, Stanford, California 94305, USA. [2]Department of Geosciences, Boise State University, Boise, Idaho 83705, USA. [3]Resource Development International – Cambodia, PO Box 494, Phnom Penh, Cambodia.

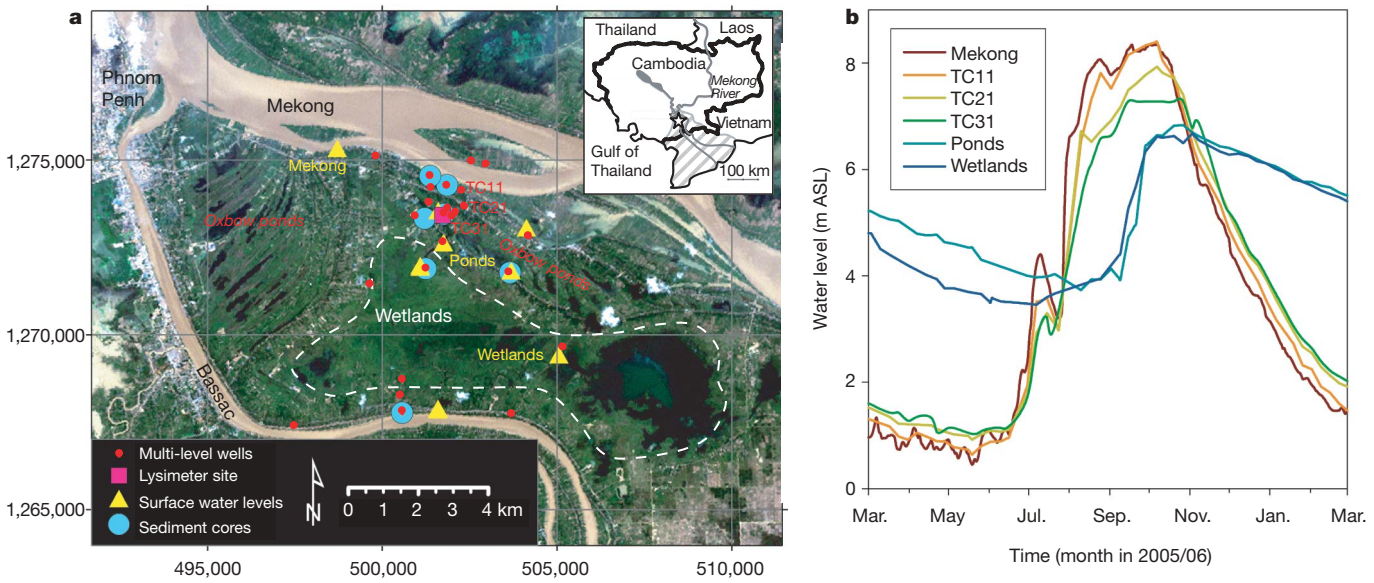


Figure 1 | Field area map and water levels. a, Our field area is located in the upper Mekong delta of Cambodia. Each multi-level well symbol represents 3–5 wells sampled at varying depths. The base map[29] is a 'true colour' composite of a Landsat image taken on 11 July 2001; the white dashed line shows the areal extent of the central wetlands, and abandoned river channel ponds are indicated. Axis tick marks represent metres in Universal Transverse Mercator zone 48N. **b**, Year-long hydrographs for selected wells (TC11, TC21 and TC31) and surface water level (Mekong River, abandoned river channel ponds, and wetlands) monitoring sites, labelled in **a**. ASL, above sea level.

and transport through the aquifer are in approximate balance with depositional delivery.

Aqueous- and solid-phase concentration profiles, chemical gradients, biogeochemical signatures and groundwater flow paths indicate that a large fraction of the arsenic entering solution is released from solids via reductive processes in near-surface soils/sediments. Within the aqueous phase, there is a steep gradient in arsenic concentrations downwards in near-surface soil/clay sediments, from $<10\,\mu g\,l^{-1}$ at the surface water–soil interface, to $\sim 600\,\mu g\,l^{-1}$ at a depth of 4 m below the surface in old river channel ponds (Fig. 2) and $>900\,\mu g\,l^{-1}$ within shallow pore water below a permanently saturated region of the wetlands. This increase in dissolved arsenic within the upper soil/sediment profile is mirrored by sharp changes in solid-phase arsenic concentrations, with $\sim 12\,mg\,kg^{-1}$ arsenic concentrations in the youngest sediments near the water table, decreasing to $<4\,mg\,kg^{-1}$ in older, permanently saturated deeper clays. Within the aquifer, solid-phase arsenic concentrations are lower (Fig. 2) and show little variation with depth, similar to Bangladesh aquifer sands[23].

The highest near-surface concentrations of dissolved arsenic occur in topographically low areas, as illustrated by old river channel ponds

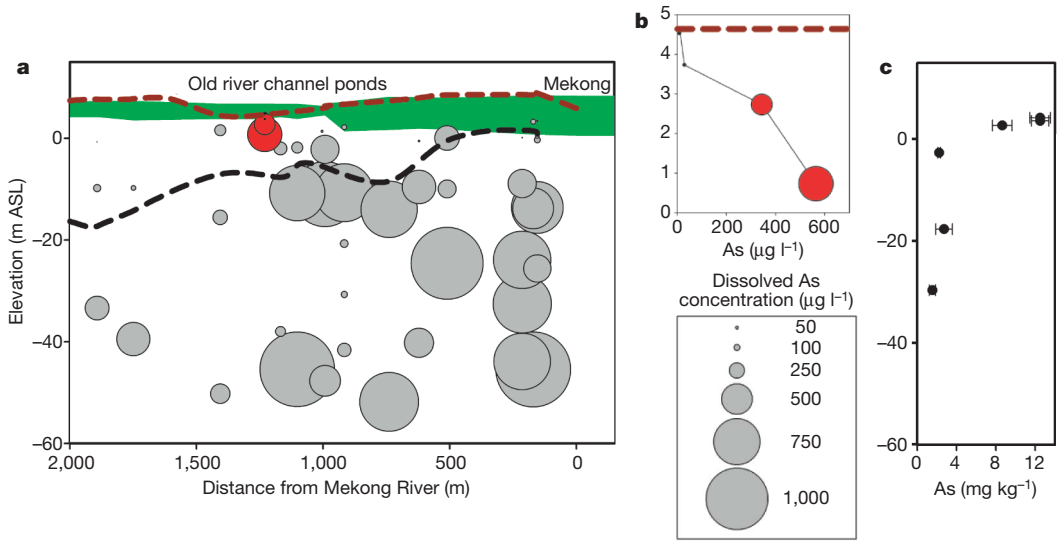


Figure 2 | Dissolved and solid-phase arsenic profiles throughout the field area. a, Field area cross-section, showing groundwater arsenic concentrations (As, grey and red filled circles; concentrations are proportional to symbol size, see key). Green, zone of variable saturation; red dashed line, ground surface; black dashed line, clay/silt-sand transition. Well nest distances are normalized based on relative perpendicular distances from ponds and Mekong River. **b**, Lysimeter arsenic concentrations (red filled circles, see key for meaning of symbol size), showing increasing arsenic with depth at the near-surface during downward flow conditions. Data taken on 21 December 2005. **c**, Solid-phase arsenic concentrations. The highest values are found within the uppermost clay sediments; within the aquifer sands, values average $2.8\,mg\,kg^{-1}$ (standard deviation, $1.65\,mg\,kg^{-1}$). Error bars, s.d. of replicate measurements.

(Fig. 2), where recently high rates of sedimentary deposition are coupled with long periods of water inundation and above-average labile carbon delivery. Persistent reducing conditions result in Fe(III) and As(V) reduction (for example, As(III) exists within both aqueous and solid phases (Supplementary Fig. 9)), and concomitant average solid-phase arsenic depletion of $\sim$0.7 mg kg^{-1} m^{-1} over the initial 14 m of the flow path below the ground surface. Correspondingly, dissolved arsenic concentrations increase by $\sim$150 μg l^{-1} m^{-1} along the initial 4 m of the flow path and then increase further by $\sim$20 μg l^{-1} m^{-1} through the remaining clay to the aquifer sands, values over two orders of magnitude greater than those along deep aquifer flow lines. Organic carbon oxidation rates, as measured by dissolved inorganic carbon (DIC) concentrations, further support arsenic liberation via near-surface anaerobic microbial respiration; DIC concentrations increase from 50 mg l^{-1} in surface waters to 300 mg l^{-1} at 5–10 m depth. Moreover, DIC is much younger ($<$1,800 yr, values consistent with similar measurements in Bangladesh aquifers[3]) than the aquifer and must therefore be primarily derived from more recent (that is, near-surface) organic carbon sources.

Temporal variations in arsenic concentrations derived from seasonal fluctuations are limited to discharge (proximate to river) and recharge (shallow sediment) zones, highlighting the mobility of arsenic across the reduced aquifer (Supplementary Fig. 8). Recharge zone porewater arsenic concentrations are highest when steep downward gradients induce water flow into the underlying aquifer, indicative of reductive arsenic release from surficial sediments during seasonal saturation. Shallow pore waters containing maximum arsenic values are depleted of oxygen and sulphate and contain measurable dissolved ferrous iron and ammonium.

Arsenic undergoes desorption from solids upon Fe(III) and As(V) reduction. Within basins of southeast Asia, it follows that arsenic release from Himalayan-derived sediments will be initiated at the point, in time and space, corresponding to the aerobic–anaerobic transition. Rates of release will initially be high, owing to the large pool of available solid-phase arsenic, and release will continue until either arsenic is depleted from solids or reduction becomes limited (owing to, for example, labile organic carbon depletion). Recent observations of arsenic release in laboratory-manipulated experiments with sediments from the oxic–anoxic interface in West Bengal[18] and increasing arsenic with groundwater age over the upper 20 m of a Bangladesh aquifer[24] support this model. Within our Mekong delta field area, we note transitions to anaerobic conditions and ensuing iron and arsenic reduction at or near the water table—where we also observe the steepest gradients in arsenic release to the pore water, as similarly noted for an arsenic-contaminated Red River delta site[25].

Observed near-surface sources of arsenic neither preclude, nor necessarily conflict with, continued arsenic release at depth through native or introduced carbon sources[3,26–28]. However, arsenic release from older, deeper aquifer sediments will most probably occur at much slower rates than from fresh near-surface wetland sediments, as evidenced by relative solid and aqueous geochemical gradients. Thus, high concentrations observed in older waters at the base of the aquifer (Figs 2, 3), particularly those more distant from the Mekong River, may reflect slow release from aquifer sediments at depth, as suggested previously[8,16,17], or a receding plume from a (potentially historic) high release zone located within the wetlands at the flow path origin.

Although the processes conspiring to produce the observed arsenic profiles are complex, it is evident that hydrology-driven arsenic release and transport from near-surface sediments represents an appreciable—and potentially dominant—source of arsenic to the aquifer, a finding with important implications for management of the arsenic problem. Moreover, these results suggest a shift in the appropriate model from dominant geochemical control towards substantial hydrologic control of arsenic in southeast Asian ground water.

Because groundwater pumping is minimal and large areas remain undeveloped and uncultivated at our field site, observed hydrologic and geochemical gradients are naturally derived and contrast with local gradients dominated by groundwater pumping, such as those in Bangladesh. Therefore, at our field site, arsenic cycling pre-dates human influence, and our results provide a potential baseline model for understanding modern arsenic contamination where land use changes have transpired. Aquifer arsenic concentrations are controlled in part by biogeochemical release from near-surface sediments and hydrologic transport, processes that at present combine to deliver arsenic at levels comparable to its efflux from the aquifer. As a result, impending or continuing changes (including upstream damming, changing land use, increased irrigation and clay excavation) that disrupt the hydrologic regime, associated biogeochemical conditions, or arsenic source material will have potentially significant consequences for arsenic concentrations in the aquifer. Although the specific impact of human activities on arsenic concentrations will be

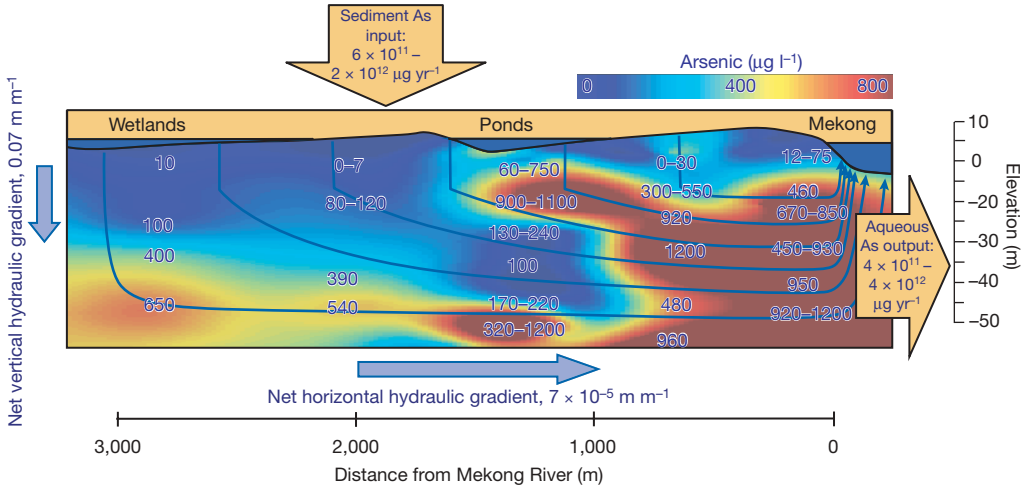


Figure 3 | Field area cross-section with groundwater flow paths and arsenic concentrations. Well water arsenic concentrations are depicted by the numbers within the cross-section, and these are contoured by kriging; temporal variations within wells are averaged, and ranges are representative of wells of equivalent depths and locations. Modelled net annual groundwater flow lines are depicted by blue arrows within the cross-section; net annual vertical gradients are 0.07 m m^{-1} in the downward direction and net annual horizontal gradients are 7×10^{-5} m m^{-1} towards the river. Arsenic inputs to the field area via sedimentation are approximately equivalent to arsenic outputs via groundwater discharge.

influenced by local site conditions, the coupling of regional hydrology with arsenic behaviour provides a framework for understanding and predicting current and future groundwater quality.

METHODS SUMMARY

Our field area is located in Kien Svay District, Kandal Province, Cambodia, in the upper reaches of the Mekong River delta. Wells were installed using a locally developed, manually driven, direct rotary method, and were backfilled with sand and native clay; lysimeters were installed into auger-dug holes and backfilled with native clay. The majority of wells were sampled once in the dry season and once in the wet season, and a subset of wells were sampled >4 times throughout the year. Lysimeters were sampled approximately twice a month. Water samples were analysed by standard methods. Intact sediment samples were obtained during well drilling, preserved in anaerobic pouches, and stored at 4 °C. Water levels in wells were measured weekly using an electronic measuring tape and surface water levels were measured weekly using a weighted measuring tape from points of fixed height. All water levels were calibrated to the Mekong River Commission stage levels following elevation surveying. Aquifer parameters were established by slug tests, constant-head permeameter tests, particle size analyses and daily tidal monitoring.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 22 May 2007; accepted 15 May 2008.

1. Smith, A. H., Lingas, E. O. & Rahman, M. Contamination of drinking-water by arsenic in Bangladesh: A public health emergency. *Bull. World Health Organ.* **78,** 1093–1103 (2000).
2. Yu, W. H., Harvey, C. M. & Harvey, C. F. Arsenic in groundwater in Bangladesh: A geostatistical and epidemiological framework for evaluating health effects and potential remedies. *Wat. Resour. Res.* 39, art no. 1146 (2003).
3. Harvey, C. F. *et al.* Arsenic mobility and groundwater extraction in Bangladesh. *Science* **298,** 1602–1606 (2002).
4. Aggrawal, P. K., Basu, A. R. & Julkarni, K. M. Comment on 'Arsenic mobility and groundwater extraction in Bangladesh'. *Science* **300,** 584b (2003).
5. van Geen, A., Zheng, Y., Stute, M. & Ahmed, K. M. Comment on 'Arsenic mobility and groundwater extraction in Bangladesh'. *Science* **300,** 584c (2003).
6. Harvey, C. F. *et al.* Response to comments on 'Arsenic mobility and groundwater extraction in Bangladesh'. *Science* **300,** 584d (2003).
7. van Geen, A. *et al.* Spatial variability of arsenic in 6000 tubewells in a 25 km^2 area of Bangladesh. *Wat. Resour. Res.* 39, art no. 1140 (2003).
8. McArthur, J. M. *et al.* Natural organic matter in sedimentary basins and its relation to arsenic in anoxic groundwater: The example of West Bengal and its worldwide implications. *Appl. Geochem.* **19,** 1255–1293 (2004).
9. Polizzotto, M. L., Harvey, C. F., Sutton, S. R. & Fendorf, S. Processes conducive to the release and transport of arsenic into aquifers of Bangladesh. *Proc. Natl Acad. Sci. USA* **102,** 18819–18823 (2005).
10. Harvey, C. F. *et al.* Groundwater dynamics and arsenic contamination in Bangladesh. *Chem. Geol.* **228,** 112–136 (2006).
11. Berg, M. *et al.* Arsenic contamination of ground and drinking water in Vietnam: A human health threat. *Environ. Sci. Technol.* **35,** 2621–2626 (2001).
12. Japan International Cooperation Agency. *The Study on Groundwater Development in Southern Cambodia* (Kokusai Kogyo Co., Tokyo, Japan, 2002).
13. Stanger, G., Truong, T. V., Ngoc, K. S. L. T. M., Luyen, T. V. & Thanh, T. T. Arsenic in groundwaters of the Lower Mekong. *Environ. Geochem. Health* **27,** 341–357 (2005).
14. Polya, D. A. *et al.* Arsenic hazard in shallow Cambodian groundwaters. *Mineral. Mag.* **69,** 807–823 (2005).
15. Berg, M. *et al.* Magnitude of arsenic pollution in the Mekong and Red River Deltas – Cambodia and Vietnam. *Sci. Tot. Environ.* **372,** 413–425 (2007).
16. Nickson, R. *et al.* Arsenic poisoning of Bangladesh groundwater. *Nature* **395,** 338 (1998).
17. Smedley, P. L. & Kinnibugh, D. G. A review of the source, behaviour, and distribution of arsenic in natural waters. *Appl. Geochem.* **17,** 517–568 (2002).
18. Islam, F. S. *et al.* Role of metal-reducing bacteria in arsenic release from Bengal delta sediments. *Nature* **430,** 68–71 (2004).
19. Klump, S. *et al.* Groundwater dynamics and arsenic mobilization in Bangladesh assessed using noble gases and tritium. *Environ. Sci. Technol.* **40,** 243–250 (2006).
20. Ta, T. K. O. *et al.* Holocene delta evolution and sediment discharge of the Mekong River, Southern Vietnam. *Quat. Sci. Rev.* **21,** 1807–1819 (2002).
21. Nguyen, V. L., Ta, T. K. O. & Tateishi, M. Late Holocene depositional environments and coastal evolution of the Mekong River Delta, Southern Vietnam. *J. Asian Earth Sci.* **18,** 427–439 (2000).
22. Tamura, T. *et al.* Depositional facies and radiocarbon ages of a drill core from the Mekong River lowland near Phnom Penh, Cambodia: Evidence for tidal sedimentation at the time of Holocene maximum flooding. *J. Asian Earth Sci.* **29,** 585–592 (2007).
23. Swartz, C. H. *et al.* Mobility of arsenic in a Bangladesh aquifer: Inferences from geochemical profiles, leaching data, and mineralogical characterization. *Geochim. Cosmochim. Acta* **68,** 4539–4557 (2004).
24. Stute, M. *et al.* Hydrological control of As concentrations in Bangladesh groundwater. *Wat. Resour. Res.* (in the press).
25. Postma, D. K. *et al.* Arsenic in groundwater of the Red River floodplain, Vietnam: Controlling geochemical processes and reactive transport modeling. *Geochim. Cosmochim. Acta* **71,** 5054–5071 (2007).
26. Rowland, H. A. L. *et al.* The control of organic matter on microbially mediated iron reduction and arsenic release in shallow alluvial aquifers, Cambodia. *Geobiology* **5,** 281–292 (2007).
27. Lear, G., Song, B., Gault, A. G., Polya, D. A. & Lloyd, J. R. Molecular analysis of arsenate-reducing bacteria within Cambodian sediments following amendment with acetate. *Appl. Environ. Microbiol.* **73,** 1041–1048 (2007).
28. Pederick, R. L., Gault, A. G., Charnock, J. M., Polya, D. A. & Lloyd, J. R. Probing the biogeochemistry of arsenic: Response of two contrasting aquifer sediments from Cambodia to stimulation by arsenate and ferric iron. *J. Environ. Sci. Health A* **42,** 1763–1774 (2007).
29. Landsat map ID 039-880, 11 July 2001 (Global Land Cover Facility, University of Maryland); downloaded for Path 126, Row 052 using the Earth Science Data Interface ⟨http://glcfapp.umiacs.umd.edu:8080/esdi/index.jsp⟩.

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by Stanford University, US NSF and US EPA STAR. We thank K. Ouch, K. Phan, co-workers at Resource Development International, G. Li, M. Meyer, M. Busbee and A. Aziz for field and laboratory assistance; S. Ganguly for modelling assistance; and A. Boucher for help with spatial analyses.

Author Contributions All authors contributed to the intellectual design, execution, interpretation and analyses presented in this study. M.L.P. assessed groundwater hydrology and geochemistry; B.D.K. analysed near-surface biogeochemistry; S.G.B. established the hydrologic framework (field layout and data collection) and conducted the modelling; M.S. facilitated the field work and provided the scientific history of the area; S.F. provided the project impetus, biogeochemical deduction, and, with S.G.B. and M.L.P., site selection. M.L.P., S.G.B. and S.F. wrote the manuscript with input from B.D.K.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to S.F. (fendorf@stanford.edu).

METHODS

Well installation, sediment collection and lysimeter installation. Wells were installed using a local drilling method by manually rotating a 1.5-inch-diameter pipe with a 4-inch-diameter open cutting tip. Lengths of 3 m were added sequentially to extend the pipe to desired well depths (up to 60 m), and water was pumped downward through the middle to physically displace the sediment. Once drilling was complete, the pipe was removed from the hole and 1.25-inch-diameter PVC tubing was installed to create the wells. At each location, 3–5 wells were put in, spanning the following depths: shallow (8–12 m), medium (20–30 m) and deep (36–60 m). Discrete, pre-fabricated well screens were used at the bottom of the PVC; screening intervals were 6–8 m for shallow wells and 4 m for medium and deep wells. Once the PVC was installed, holes were backfilled with coarse sand and capped with clay and/or cement.

Because of the potential for homogenization and oxidation of drill cuttings, an alternative, intact coring procedure was used at selected locations for sediment retrieval. A 0.75-inch open core device fitted with a polycarbonate sleeve and core-catcher was deployed through the drilling pipe at 3 m intervals and driven into the undisturbed sediment below the drilling tip. Once retrieved, samples were immediately capped and sealed in O_2-impermeable pouches with AnaeroPacks (Mitsubishi Gas Chemical) to prevent oxidation of the sediments. Sediment samples were stored and transported at 4 °C.

In order to monitor near-surface porewater chemistry at discrete depths, ceramic cup lysimeters were installed in a sediment profile in the abandoned river channel ponds at depths of 0.1, 0.5, 1, 2 and 4 m. Holes were dug with a hand auger and soil cores were collected in undisturbed sediments below the extent of augering; samples were preserved as above. Following lysimeter installation, holes were backfilled with native clay.

Clay sediment samples for ^{14}C analysis were obtained at multiple depths from a ~16 m pit created by a mechanical excavator. Samples were collected from the pit wall using 0.5 m copper pipes that were pounded horizontally into the sediments. The pipes were dug out and ends were immediately sealed in the field with paraffin wax. The central portions of the cores were used for ^{14}C analyses conducted at the NSF-University of Arizona AMS facility.

Water level measurements. Hydraulic heads in each well were measured weekly with an electronic water level tape. Surface water levels were also measured weekly with a weighted measuring tape from points of fixed height. All reference points for water level measurements were spatially linked by vertical surveying with an auto level (manufacturer reported 2.5 mm accuracy per double km run); field error was <5 mm on closed loops.

Elevations were referenced with Mekong River stage levels provided by the Mekong River Commission. Distance-weighted averages from the Phnom Penh port (upstream of our site) and the Neak Luong (downstream) stations were used to calibrate the absolute elevation of our water level monitoring site in the Mekong River, and the remainder of the water level measurements throughout our field area were adjusted accordingly.

Water sampling. Wells were sampled with a peristaltic pump at flow rates of ~1 l min^{-1}. A multiparameter probe equipped with a flow-through cell was placed in the outflow line to monitor dissolved oxygen, pH, conductivity, temperature and E_h. Wells were purged before sample collection until multiparameter stabilization was observed (typically 30–130 min). If a well could not be pumped continuously (that is, low yield), it was pumped dry and the re-infiltrated water was collected on the following day.

Groundwater samples were filtered with 0.45 μm filters and collected in acid-washed bottles. Samples for cations and total arsenic concentrations were acidified with trace-metal grade HCl to pH < 2. Samples for anion analyses were pretreated with a Bio-Rad AG50W-X8 cation exchange resin in hydrogen form (Bio-Rad Laboratories) to prevent oxidative metal precipitation and subsequent anion scavenging. Arsenic speciation in the field was performed by acidification of ground water to pH 3, followed by treatment with a Bio-Rad AG1-X8 anion exchange resin in acetate form to remove As(v). Dissolved organic carbon (DOC), nitrate and ammonium samples were acidified with HCl and sterilized with $HgCl_2$. The majority of wells were sampled once during the wet season and once during the dry season. A subset of wells along a transect from the Mekong River was sampled approximately monthly.

Lysimeters were allowed to equilibrate with pore water for 30 d before sample collection; during this time, samples were routinely collected and discarded. Each lysimeter was subsequently sampled bimonthly from August 2005 to September 2006. Pore water was drawn by suction (−800 mbar) into bottles, each containing 15 ml of trace-metal grade 3 M HCl for preservation of As and Fe.

Surface water was collected in a churn bucket at the water–air interface and sampling was performed as with the groundwater samples. Surface water samples were obtained monthly.

Analytical measurements. Measurements on aqueous samples were conducted in the field and in the laboratory. Field measurements were performed for arsenic, alkalinity, ferrous iron, nitrate, sulphate and sulphide; all but alkalinity and dissolved sulphide were duplicated in the laboratory. Alkalinity measurements were performed by pH-verified colorimetric titration. Ferrous iron, nitrate, sulphate and sulphide were measured using standard colorimetric spectroscopic methods.

In the laboratory, aqueous-phase arsenic concentrations were analysed by hydride generation inductively-coupled-plasma atomic emission spectroscopy (HG-ICP-AES). Arsenate was reduced to arsenite with KI, and arsine gas was subsequently formed by reaction with a 0.6% $NaBH_4$/0.5% NaOH solution. Detection limits were 5 μg l^{-1} arsenic. ^{14}C measurements of DIC were performed at the University of Waterloo Environmental Isotopes Laboratory. Elemental solid-phase concentrations were determined after microwave digestion. Sediment samples were chemically dissolved in 3:1 concentrated HNO_3:concentrated HF and the resulting solution was evaporated to dryness. Following reconstitution in HCl, concentrations were measured by ICP-AES.

LETTERS

Stress changes from the 2008 Wenchuan earthquake and increased hazard in the Sichuan basin

Tom Parsons[1], Chen Ji[2] & Eric Kirby[3]

On 12 May 2008, the devastating magnitude 7.9 (Wenchuan) earthquake struck the eastern edge of the Tibetan plateau, collapsing buildings and killing thousands in major cities aligned along the western Sichuan basin in China. After such a large-magnitude earthquake, rearrangement of stresses in the crust commonly leads to subsequent damaging earthquakes[1–5]. The mainshock of the 12 May earthquake ruptured with as much as 9 m of slip along the boundary between the Longmen Shan and Sichuan basin, and demonstrated the complex strike–slip and thrust motion[6] that characterizes the region[7,8]. The Sichuan basin and surroundings are also crossed by other active strike–slip and thrust faults. Here we present calculations of the coseismic stress changes that resulted from the 12 May event using models of those faults, and show that many indicate significant stress increases. Rapid mapping of such stress changes can help to locate fault sections with relatively higher odds of producing large aftershocks.

Globally, earthquakes like the magnitude (M) 7.9 shock that struck the western Sichuan region can be associated with triggered aftershocks[5]

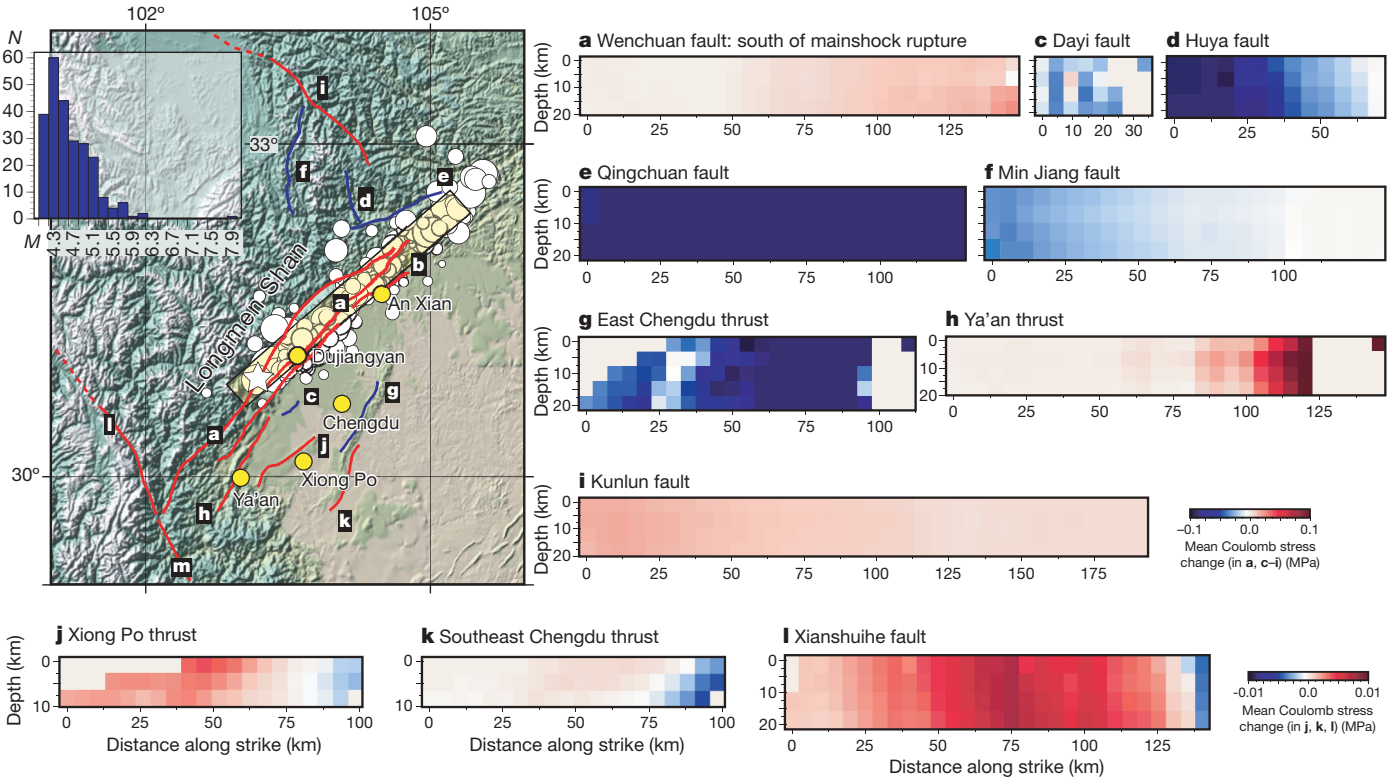


Figure 1 | Map of study area, calculated stress changes on major Sichuan basin and other faults after the Wenchuan earthquake, and aftershock data. Faults are labelled **a**–**m** in the map, and panels displaying calculated stress changes for most of these faults are shown; the zero of distance on each x axis marks the south end of the fault. Stress increases (>0.01 MPa) have been demonstrated to bring faults to failure, with delays ranging from seconds to decades[2,5,12,13]. The epicentre is marked with a white star, aftershocks ($M \geq 4$) are shown as white circles, and the rupture plane is shown as an outlined yellow rectangle and coincides with **a**, the Beichuan and Wenchuan faults. The Wenchuan fault extends south of the rupture, and shows a calculated stress increase (data panel **a**). Whether **b**, the Pengguan fault, was involved as well is uncertain, so no stress change calculations are shown for it. Most modelled faults show stress increases (red colours) except for **c**, Dayi, **d**, Huya, **e**, Qingchuan, **f**, Min Jiang, and **g**, the thrust east of Chengdu (blue shading). Note that the colour scales differ by region. The **h**, Ya'an thrust, **i**, Kunlun fault, **j**, Xiong Po thrust, **k**, thrust southeast of Chengdu, and **l**, the Xianshuihe fault all have calculated stress increases. Calculated stress changes on **m**, the Shimian fault, were negligible, so no data panel is shown for it. Yellow circles on the map indicate significant population centres. Map inset, aftershock magnitude–frequency (M–N) distribution as of 5 June 2008.

[1]US Geological Survey, MS-999, 345 Middlefield Road, Menlo Park, California 94025, USA. [2]Department of Geological Sciences, University of California, Santa Barbara, California 93106, USA. [3]Department of Geosciences, Penn State University, University Park, Pennsylvania 16802, USA.

with $M > 7$; a recent example that affected a large population is the 1999 Izmit ($M = 7.4$) and Düzce ($M = 7.1$) pairing in Turkey. Stress-transfer analysis of those events came too late to be any factor for mitigation[4,9], but the technique was more successful after the great 2004 Sumatra earthquake, when a $M = 8.7$ shock struck three months later in a region calculated[3] to have been stressed by the mainshock. For the 12 May 2008 event, early-stage calculations of coseismic stress transfer onto Sichuan basin faults are made by stepping through broad parameter ranges because exact values remain unknown. This approach enables rapid mapping of faults with heightened rupture likelihood (Fig. 1), and allows an opportunity for prospective testing of static stress transfer effects on earthquake hazard.

The 12 May $M = 7.9$ earthquake appears to have ruptured the Beichuan fault (labelled, together with the Wenchuan fault, as **a** in the map in Fig. 1) along the edge of the Longmen Shan (Fig. 1). The right-lateral oblique Pengguan fault[7,8] (Fig. 1b) parallels the Beichuan fault, and it is unclear at the time of this writing whether this fault was directly involved in the mainshock rupture; if it was not involved, meaningful stress change calculations cannot be made on its surface because it is so close to the rupture[10]. Coulomb failure stress is calculated to have been increased south of the mainshock rupture on the Wenchuan fault (Fig. 1a) by up to 0.1 MPa. The active right-lateral[8] strike–slip Dayi fault parallels the mainshock rupture, and as a result, is calculated to have reduced Coulomb stress (Fig. 1c). Thrust faults northwest of the rupture include the Huya, Qingchuan and Min Jiang zones[11] (Fig. 1d–f), and all indicate Coulomb stress reduction.

An array of thrust faults underlies the Sichuan basin, paralleling the range front (Fig. 1). These faults change in character from west to east across the basin; close to the range front, they dip to the northwest and appear to accommodate some right-lateral slip, whereas on the eastern side of the basin, faults dip shallowly down to the east[7,8]. The thrust fault near Ya'an (Fig. 1h) lies south of the mainshock rupture and thus is calculated to have increased Coulomb stress. Similarly, the thrust faults bounding the Sichuan basin southeast of Chengdu (Fig. 1k) and near Xiong Po (Fig. 1j) also are calculated to have increased stress. Increases are calculated to be much smaller on the east side of the basin ($\sim$0.01 MPa) because they are located farther from the mainshock. The thrust fault immediately east of Chengdu (Fig. 1g) has a calculated stress decrease (about -0.01 MPa).

The Longmen Shan block is bound by left-lateral strike–slip faults to the north and south. The left-lateral Xianshuihe fault strikes northwest of its junction with the Wenchuan fault and has a calculated Coulomb stress increase over a 125-km length from the junction (Fig. 1l). South of the junction, left-lateral motion is taken up by the Shimian fault[7] (Fig. 1m). Coulomb calculations on this segment show negligible changes. Similarly, calculations on the left-lateral Qinling fault, northeast of the Longmen Shan region, indicate negligible changes. However, the left-lateral Kunlun fault (Fig. 1i) northwest of the rupture shows a calculated stress increase. Thus it appears that most faults with calculated stress increases are confined to the southern Sichuan basin and boundaries, with the exception of the left-lateral Kunlun and Xianshuihe faults north and south of the basin.

The first 25-day period of aftershock activity ($M \geq 4$) is mostly confined within the mainshock rupture area (Fig. 1). The largest aftershocks within the period were two $M = 6$ events; the aftershock magnitude–frequency distribution (Fig. 1) suggests a potential moment deficiency in the $M = 5$–6.5 range, although variability is high and the number of events is small. Globally, static stress changes from $M > 7$ earthquakes are correlated with seismicity-rate changes over distances in excess of 200–250 km from the mainshock[5].

The 12 May 2008 $M = 7.9$ earthquake that struck the eastern Sichuan region caused grievous losses, yet its legacy includes possible large aftershocks in the near future because it increased failure stress on important faults within and around the Sichuan basin. Given that delays of years to decades between mainshocks and large aftershocks are commonly observed around the world[2,4], identifying potential future rupture zones will be useful in focusing mitigation efforts.

Received 16 May; accepted 19 June 2008.
Published online 6 July 2008.

1. Stein, R. S., Barka, A. A. & Dieterich, J. H. Progressive failure on the North Anatolian fault since 1939 by earthquake static stress triggering. *Geophys. J. Int.* **128,** 594–604 (1997).
2. Stein, R. S. The role of stress transfer in earthquake occurrence. *Nature* **402,** 605–609 (1999).
3. McCloskey, J., Nalbant, S. S. & Steacy, S. Indonesian earthquake: Earthquake risk from co-seismic stress. *Nature* **434,** 291 (2005).
4. Parsons, T., Toda, S., Stein, R. S., Barka, A. & Dieterich, J. H. Heightened odds of large earthquakes near Istanbul: An interaction-based probability calculation. *Science* **288,** 661–665 (2000).
5. Parsons, T. Global Omori law decay of triggered earthquakes: Large aftershocks outside the classical aftershock zone. *J. Geophys. Res.* **107,** doi:10.1029/2001JB000646 (2002).
6. Ji, C. & Hayes, G. Preliminary result of the May 12, 2008 Mw 7.9 eastern Sichuan, China earthquake. ⟨http://earthquake.usgs.gov/eqcenter/eqinthenews/2008/us2008ryan/finite_fault.php⟩ (2008).
7. Burchfiel, B. C., Chen, Z., Liu, Y. & Royden, L. H. Tectonics of the Longmen Shan and adjacent regions. *Int. Geol. Rev.* **37,** 661–735 (1995).
8. Densmore, A. L. *et al.* Active tectonics of the Beichuan and Pengguan faults at the eastern margin of the Tibetan Plateau. *Tectonics* **26,** doi:10.1029/2006TC001987 (2007).
9. Hubert-Ferrari, A. *et al.* Seismic hazard in the Marmara Sea region following the 17 August 1999 Izmit earthquake. *Nature* **404,** 269–273 (2000).
10. Steacy, S., Marsan, D., Nalbant, S. S. & McCloskey, J. Sensitivity of static stress calculations to the earthquake slip distribution. *J. Geophys. Res.* **109,** doi:10.1029/2002JB002365 (2004).
11. Kirby, E. & Whipple, K. X. Distribution of active rock uplift along the eastern margin of the Tibetan Plateau: Inferences from bedrock channel longitudinal profiles. *J. Geophys. Res.* **108,** doi:10.1029/2001JB000861 (2003).
12. Harris, R. A. Introduction to special section: Stress triggers, stress shadows, and implications for seismic hazard. *J. Geophys. Res.* **103,** 24347–24358 (1998).
13. Reasenberg, P. A. & Simpson, R. W. Response of regional seismicity to the static stress change produced by the Loma Prieta earthquake. *Science* **255,** 1687–1690 (1992).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank R. Harris and W. Thatcher for their help with this manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to T.P. (tparsons@usgs.gov).

LETTERS

Subtropical to boreal convergence of tree-leaf temperatures

Brent R. Helliker[1] & Suzanna L. Richter[2]

The oxygen isotope ratio ($\delta^{18}O$) of cellulose is thought to provide a record of ambient temperature and relative humidity during periods of carbon assimilation[1,2]. Here we introduce a method to resolve tree-canopy leaf temperature with the use of $\delta^{18}O$ of cellulose in 39 tree species. We show a remarkably constant leaf temperature of 21.4 ± 2.2 °C across 50° of latitude, from subtropical to boreal biomes. This means that when carbon assimilation is maximal, the physiological and morphological properties of tree branches serve to raise leaf temperature above air temperature to a much greater extent in more northern latitudes. A main assumption underlying the use of $\delta^{18}O$ to reconstruct climate history is that the temperature and relative humidity of an actively photosynthesizing leaf are the same as those of the surrounding air[3,4]. Our data are contrary to that assumption and show that plant physiological ecology must be considered when reconstructing climate through isotope analysis. Furthermore, our results may explain why climate has only a modest effect on leaf economic traits[5] in general.

The ratio of stable oxygen isotopes ($\delta^{18}O$) in tree-ring cellulose was first used to reconstruct temperatures during tree growth, and a seminal study[6] showed a strong correlation between $\delta^{18}O$ of woody tissue and mean annual temperature (MAT).

Temperature is the main controlling factor for the isotopic composition of precipitation[7], and the $\delta^{18}O$ of precipitation is the primary control on the $\delta^{18}O$ of tree-ring cellulose. Methodological advancements have enabled the measurement of $\delta^{18}O$ in smaller increments of cellulose and have shown that inter-annual and intra-annual variation in tree-ring $\delta^{18}O$ can correlate with episodic or cyclic events, such as tropical cyclones and the El Niño Southern Oscillation, that cause a wholesale change in the $\delta^{18}O$ of precipitation water[8,9].

A secondary control on the $\delta^{18}O$ of tree-ring cellulose is relative humidity through its effects on evaporation. A greater evaporative gradient (lower relative humidity) results in greater loss of the lighter isotope (^{16}O), thereby enriching leaf water in the heavy isotope ^{18}O; this leaf-water isotopic signal is incorporated into sucrose and ultimately into cellulose. Accordingly, several studies have shown that the $\delta^{18}O$ of sucrose, and consequently tree-ring cellulose, records ambient relative humidity[2–4,10,11] and that the recorded humidity signal is more strongly associated with periods of maximal photosynthesis[12,13]. Hence, analysis of the $\delta^{18}O$ of tree-ring cellulose offers the potential to reconstruct the year-to-year history of growth temperature and relative humidity in both living and subfossil trees.

The historical assumption in isotope studies of tree rings has been that leaves in the tree canopy were coupled to the ambient environment, and thus that the temperature and relative humidity experienced by the leaf were equal to those of ambient air. However, it has long been established that leaf temperatures can deviate from ambient temperatures because variation in water loss, convective heat loss, and reflectance can increase or decrease leaf temperatures[14,15]. Leaf temperature affects the evaporative gradient from leaf to air because temperature exponentially affects the saturation vapour pressure of water vapour inside the leaf. Hence, any adaptation that leads to a systematic offset between leaf and ambient temperature can cause leaf relative humidity to be markedly different from ambient relative humidity. The effects of leaf morphology, temperature, and water loss on the enrichment of ^{18}O in leaf water have been well documented at the leaf level in a variety of plants[16–19], but these effects have been largely ignored in tree-ring isotope work.

Whereas the $\delta^{18}O$ analysis of individual tree rings permits the reconstruction of year-to-year changes in weather, we proposed that the analysis of entire, homogenized tree rings provide a multi-year or lifespan-integrated measure of tree responses to average climate. We examined α-cellulose $\delta^{18}O$ values from the wood of 39 deciduous and evergreen tree species across 50° of latitude in North America. These samples came from a larger survey of tree-cellulose $\delta^{18}O$ to determine whether the current understanding of tree-cellulose $\delta^{18}O$ formation could explain the increasing offset between cellulose $\delta^{18}O$ and precipitation $\delta^{18}O$ as the mean annual temperature decreased[20]. Expressing observed cellulose ^{18}O as an ^{18}O enrichment above source water ($\Delta^{18}O_c$) removes the well-established temperature effects on the $\delta^{18}O$ of precipitation (and consequently cellulose $\delta^{18}O$) and highlights the distinct biological effects on cellulose $\delta^{18}O$. We found a highly significant and unexpected correlation with MAT, showing that the observed values of cellulose $\Delta^{18}O$ became more and more enriched in ^{18}O above source water as MAT decreased (filled circles in Fig. 1; $F = 844.7$, $P < 0.0001$).

In an attempt to explain this observation, we used a physiological model of isotopes in cellulose, starting from the standard assumption that tree-canopy temperature and relative humidity were coupled to ambient environmental conditions (open squares in Fig. 1). A previous cellulose $\delta^{18}O$ model[19] was parameterized by using long-term climate averages (http://www.climate.weatheroffice.ec.gc.ca/climate_normals/index_1961_1990_e.html; http://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NCEP/.CPC/.GSOD/.MONTHLY/) and model estimates of $\delta^{18}O$ in precipitation[7] to develop predictions of $\Delta^{18}O_c$. To be conservative in our predictions, we assumed a 15% error in the use of precipitation $\delta^{18}O$ for tree source water and further allowed for a large range of error in the prediction of leaf-water ^{18}O enrichment and the isotopic exchange factors of the cellulose-isotope model (error bars for predictions in Fig. 1; see Methods for details). With leaf temperature set to ambient temperature, the physiological model could not account for the higher enrichment in observed $\Delta^{18}O_c$ at lower MAT even when considering the range of error in the model (the slope was not significantly different from 0; $F = 2.3160$, $P = 0.1327$). The isotopic exchange or fractionation factors associated with cellulose formation show no change with temperature and are relatively constant across species[2,12]. Although there is some evidence that the

[1]Department of Biology, [2]Department of Earth and Environmental Science, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA.

organic-water fractionation factor may change in enriched plant source water[21], this is the reverse of the pattern observed in colder climates. The most parsimonious conclusion from Fig. 1 is therefore that leaf water becomes more enriched than expected as MAT decreases, and this enrichment is recorded in $\Delta^{18}O_c$. The most likely mechanism to cause greater enrichment is that tree-leaf temperature was generally higher than ambient temperature in colder climates; the inverse must be true for leaf relative humidity.

Recognizing that the disagreement between observed and predicted values of $\Delta^{18}O_c$ in Fig. 1 are most probably due to the assumption of equality of leaf and ambient temperatures during photosynthesis, we developed a novel application in which observed tree-cellulose $\delta^{18}O$ can be used to solve for integrated tree-canopy temperature. Tree-leaf temperatures were determined by using the observed cellulose $\Delta^{18}O_c$ data and climate averages and rearranging the same isotope model[19] to solve for the saturated leaf vapour pressure that satisfied observed $\Delta^{18}O_c$. Saturated water vapour pressure has a well-quantified relationship with temperature; a given saturated vapour pressure therefore yields a unique temperature. The mean tree-leaf temperature for all 39 species across 50° of latitude was 21.4 ± 2.2 °C (grey box in Fig. 2a). We found no significant relationship between leaf temperature and MAT or between leaf temperature and growing-season temperature.

These results suggest that most tree photosynthesis occurred when leaf temperatures were about 21 °C, irrespective of latitude and average growing-season temperatures. The $\delta^{18}O$ in a tree ring inherently represents a whole-canopy integration of leaf-level processes that are weighted towards periods of maximal carbon assimilation. This effective homeostasis of leaf temperatures means that there was a significant elevation of leaf temperatures over mean growing-season temperatures in colder climates and, as a corollary, leaf temperatures that were lower than ambient temperatures in hotter, temperate climates (Fig. 2b). In the subtropical sites, tree-leaf temperatures were much closer to ambient temperatures. The relationship of leaf temperature to ambient temperature holds across closely related congeneric species as well as through the larger phylogenetic groupings of angiosperms and gymnosperms.

The logistical constraints on direct, empirical measurements of integrated tree-canopy temperature for an entire growing season mean that there are few data sets in existence with which to compare our results. However, leaf temperature measurements on individual branches of *Abies*, *Picea* and *Pinus* species in subalpine regions of Wyoming, USA, were 5–9 °C above ambient temperatures during periods of maximal photosynthesis[22]. A recent study using infrared thermal imaging of a temperate mixed forest in Switzerland showed that the temperature of dense tree canopies was 4–5 °C higher than ambient temperature[23], and that of less dense canopies was 0.3–2.7 °C higher than ambient temperature. These observations, although indirect, are in agreement with our resolved relationships between ambient and tree-canopy temperatures (Fig. 2b).

How does leaf temperature control occur? In warmer climates, leaf temperatures are lowered by evaporative cooling and mechanisms that reduce the absorptance of solar radiation such as decreased leaf angles and reflective leaf hairs[14,15]. These adaptations are counterproductive to warming a leaf in colder environments. One established mechanism for elevating tree-leaf temperatures above ambient temperatures is to increase the number of leaves on a given length of branch, thereby increasing the branch boundary layer and consequently decreasing convective heat loss from individual leaves[22,24]. This mechanism has been shown directly in a few coniferous tree branches[22] and indirectly at the canopy scale, where greater canopy density results in greater elevation of canopy temperature over ambient temperature in both needle-leaved and broad-leaved trees[23].

Trees maintain growth and reproduction over a broad climatic spectrum through an array of physiological and morphological adaptations[22,25], yet the idea that these adaptations converge towards leaf-temperature homeostasis across biomes is new. Early work on plant acclimation to prevailing temperatures focused on the temperature range over which CO_2 assimilation was optimal[25,26]. As a general rule, these photosynthetic temperature optima were found to be lower than or equal to growing-season temperatures in hot climates and higher than the ambient temperatures in cold climates. Our results explain this observation over a broad climatic range and further suggest that the overarching trend is to maintain leaves at an optimal temperature irrespective of mean climate.

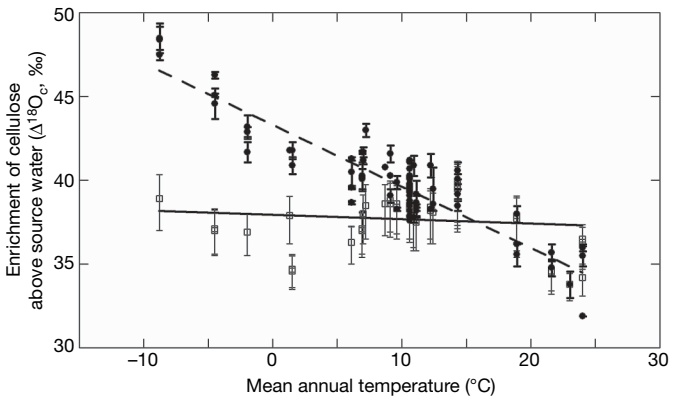


Figure 1 | Plot of predicted and observed cellulose $\Delta^{18}O_c$ against mean annual temperature. For observations of the 39 tree species at 25 sites (filled circles), each point represents a species mean and the error bars indicate the standard deviation for the observations. The predictions (open squares) were based on forcing leaf temperature to equal MAT in the isotope cellulose model[19]. The error bars for the predictions are the upper-end and lower-end predictions using the largest observed ranges in isotopic exchange factors and a ±15% error in plant source water $\delta^{18}O$. Information on species, *n* values, study sites and model estimates can be found in Methods and Supplementary Information.

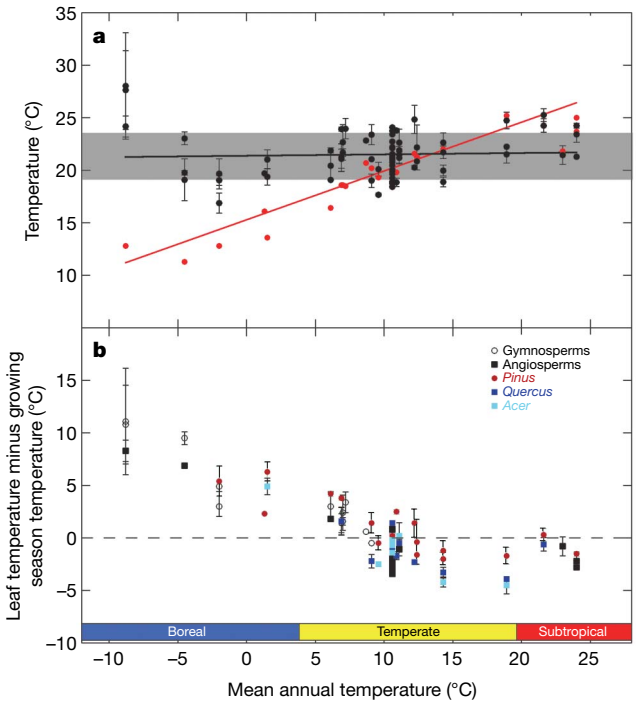


Figure 2 | Resolved tree-canopy temperature versus mean annual temperature. **a**, Leaf temperatures (black) resolved from tree-cellulose $\Delta^{18}O$, and ambient growing-season temperatures (red). The grey box indicates the mean and s.d. of the leaf temperature of all tree species (21.4 ± 2.2 °C). **b**, Plot of integrated tree-leaf temperatures minus ambient growing-season temperatures against mean annual temperature from subtropical to boreal biomes. Error bars indicate s.d.

However, we do not suggest that tree canopies maintain a constant temperature of 21 °C over the course of a day or a season. Rather, we propose that the integrated and apparent homeostatic temperature that our isotope analysis has revealed is more of a long-term target value. Over time, the morphological and physiological characteristics of a tree canopy work to maintain leaf temperatures near that target value within a climatic regime, while the functional plasticity of the photosynthetic apparatus works to maximize carbon uptake in the face of short-term variation in weather. Canopy temperatures can vary drastically in the short term with the vagaries of radiation input caused by clouds, wind speed and the diurnal change in the Sun's angle of incidence. Observed mid-day canopy temperature amplitudes can be as large as 12 °C (ref. 23) and observed temperature-response curves of photosynthesis show a similarly broad range[26]. Within species, photosynthetic temperature optima track changes in ambient temperature with the progression of the growing season. However, there are clear limits, because trees grown at temperatures higher than ambient temperature do indeed show an increase in the photosynthetic temperature optimum above trees grown at ambient conditions, but they also show a much reduced overall growth[27].

The effect of leaf temperature on the difference between leaf and ambient relative humidity has a profound effect on interpretations of $\delta^{18}O$ in plant material. Leaf relative humidity was determined by dividing the saturated vapour pressure at leaf temperature (determined by our isotope analysis) by the ambient vapour pressure from the observed mean climate data. Depending on biome—and even within biomes—the difference between ambient and leaf relative humidity can be significant (Fig. 3). In boreal systems, the differences between leaf and ambient relative humidity were nearly 30%. The broad temperature range encompassed by temperate forested systems makes generalization more difficult, but our data show that ambient relative humidity can be 5–10% above or below leaf relative humidity (Fig. 3). Our analysis shows that reconstructing ambient humidity by using tree-ring $\delta^{18}O$ becomes increasingly dubious as MAT decreases. Caution is therefore advised when interpreting tree-ring $\delta^{18}O$ data from high latitudes for both contemporary samples and samples of relictual wood from high-latitude forests of the past[28]. However, our results proffer a correction for reconstruction, and—perhaps more significantly—the new approach developed here shows that physiological responses to inter-annual temperature variation can be extracted from the $\delta^{18}O$ of tree rings if ambient climatic conditions are known.

Temperature effects on photosynthesis, respiration and water acquisition are primary factors determining tree distribution and will no doubt be of greater concern in the future as global temperatures continue to increase. However, showing specific responses to temperature is a challenging task. Here we have shown that stable-isotope analysis of the α-cellulose in whole-tree cores provides an integration of the responses of both leaf relative humidity and leaf temperature to mean climate. This basic understanding can now be applied to the more rigorous task of analysing inter-annual variation in tree ecophysiological responses to the vagaries of weather as climate has changed over the past century.

The discovery of relatively invariant leaf temperatures has two important ramifications that transcend stable-isotope studies. First, elevated canopy temperature and depressed leaf relative humidity should have a large effect on real and modelled water loss from boreal ecosystems. Second, if the architectural controls of branches on leaf temperature are as widespread as our data suggest, then direct climatic selection on the evolution of leaf traits would be relaxed, whereas the selective force of climate on other plant organs (for example stems and roots) would remain. Our results therefore offer a possible explanation for the unexpected finding[5] that climate is a minor correlate with global leaf economic traits. We propose that climatic constraints on other components of plant function, such as the avoidance of and recovery from xylem cavitation by freezing or the balance between canopy photosynthesis and stem/root respiration, could be more important for tree distribution in colder environments. Finally, the branch morphological characteristics that serve to raise leaf temperatures above ambient temperature in cold environments would inherently limit tree performance and hence distribution in warmer areas and, significantly, in areas where warming is expected to increase.

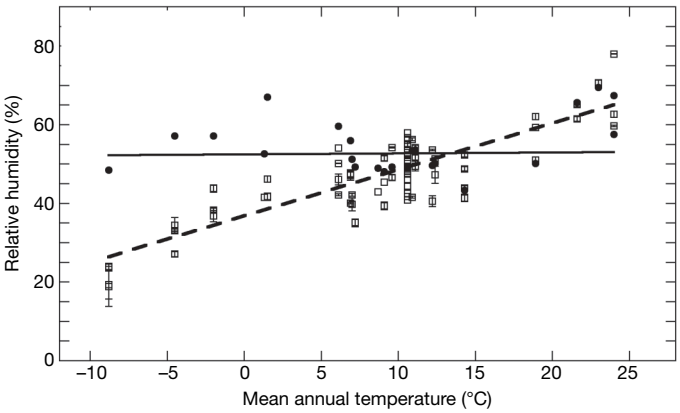


Figure 3 | Comparison of ambient and canopy-based relative humidity. Relative humidity with respect to the leaf (ambient water vapour pressure divided by saturation water vapour pressure at leaf temperature) determined from tree-ring $\Delta^{18}O$ (open squares) and ambient relative humidity (ambient water vapour pressure divided by saturation water vapour pressure at ambient temperature) during the growing season (filled circles). Error bars indicate s.d. of the resolved leaf relative humidity.

METHODS SUMMARY

Sampling and isotope analysis. Tree-ring wood was from a larger meta-analysis[20] of tree cellulose $\delta^{18}O$ (see Table S1 in Supplementary Information and Full Methods). The samples and sites were chosen on the basis of precipitation patterns and climate data availability. The wood was ground, homogenized and α-cellulose was extracted and pyrolysed at 1100 °C. The resulting CO gas was analysed in a Thermo-Finnigan Delta Plus isotope ratio mass spectrometer. All samples were run in triplicate; the standard deviation of the primary referece was 0.23‰. Isotope ratios (δ) were expressed relative to Vienna Standard Mean Ocean Water (VSMOW) by

$$\delta^{18}O = \left(\frac{R_{sample}}{R_{standard}} - 1\right) \times 1{,}000 \quad (1)$$

where R_{sample} is the number isotope ratio of $^{18}O/^{16}O$ and $R_{standard}$ (VSMOW) = 0.0020052.

Isotope discrimination above tree source water was expressed as

$$\Delta = \left(\frac{R_{sample}}{R_{source}} - 1\right) \times 1{,}000 \quad (2)$$

(approximated by $\delta^{18}O_{sample} - \delta^{18}O_{source}$), where R is the molar isotope ratio of either the cellulose or the tree source water.

Model parameterization. To develop predictions for $\Delta^{18}O_c$ we used the average of published observations for the isotope exchange values and used the range of these observations to develop a range of predictions (error bars in Fig. 1; see also Supplementary Table 2). Tree source water was assumed to be equal to precipitation-weighted model outputs[7]. For Fig. 1, leaf temperature was assumed to be equal to ambient temperature and we used observed relative humidity ($= 100 \times e_a/e_i$). Tree-canopy leaf temperature (T_L) was obtained by rearranging the isotope cellulose model to solve for the leaf saturated vapour pressure (e_i). T_L was obtained from e_i by rearranging a standard vapour–pressure–temperature relationship. For all models we used monthly mean temperature and relative humidity, weighted by net primary productivity and growing season length.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 10 March; accepted 28 April 2008.
Published online 11 June 2008.

1. Epstein, S., Thompson, P. & Yapp, C. J. Oxygen and hydrogen isotopic ratios in plant cellulose. *Science* **198**, 1209–1215 (1977).

2. Roden, J. S., Lin, G. & Ehleringer, J. R. A mechanistic model for interpretation of hydrogen and oxygen isotope ratios in tree-ring cellulose. *Geochim. Cosmochim. Acta* **64,** 21–35 (2000).
3. Anderson, W., Bernasconi, S. & McKenzie, J. Oxygen and carbon isotopic record of climatic variability in tree ring cellulose (*Picea abies*): An example from central Switzerland (1913–1995). *J. Geophys. Res.* **103,** 31,625–31,636 (1998).
4. Wright, W. E. & Leavitt, S. W. Boundary layer humidity reconstruction for a semiarid location from tree ring cellulose $\delta^{18}O$. *J. Geophys. Res. Atmos.* **111,** D18105 (2006).
5. Wright, I. J. *et al.* The worldwide leaf economics spectrum. *Nature* **428,** 821–827 (2004).
6. Gray, J. & Thompson, P. Climatic information from $^{18}O/^{16}O$ ratios of cellulose in tree rings. *Nature* **262,** 481–482 (1976).
7. Bowen, G. & Revenaugh, J. Interpolating the isotopic composition of modern meteoric precipitation. *Wat. Resour. Res.* **39,** 1299 (2003).
8. Evans, M. N. & Schrag, D. P. A stable isotope-based approach to tropical dendroclimatology. *Geochim. Cosmochim. Acta* **68,** 3295–3305 (2004).
9. Miller, D. L. *et al.* Tree-ring isotope records of tropical cyclone activity. *Proc. Natl Acad. Sci. USA* **103,** 14294–14297 (2006).
10. Robertson, I., Waterhouse, J. S., Barker, A. C., Carter, A. H. C. & Switsur, V. R. Oxygen isotope ratios of oak in east England: implications for reconstructing the isotopic composition of precipitation. *Earth Planet. Sci. Lett.* **191,** 21–31 (2001).
11. Saurer, M., Cherubini, P. & Siegwolf, R. Oxygen isotopes in tree rings of *Abies alba*: The climatic significance of interdecadal variations. *J. Geophys. Res. Atmos.* **105,** 12461–12470 (2000).
12. Cernusak, L. A., Farquhar, G. D. & Pate, J. S. Environmental and physiological controls over oxygen and carbon isotope composition of Tasmanian blue gum, *Eucalyptus globulus*. *Tree Physiol.* **25,** 129–146 (2005).
13. Gessler, A., Peuke, A. D., Keitel, C. & Farquhar, G. D. Oxygen isotope enrichment of organic matter in *Ricinus communis* during the diel course and as affected by assimilate transport. *New Phytol.* **174,** 600–613 (2007).
14. Miller, P. C. Bioclimate, leaf temperature, and primary production in red mangrove canopies in south Florida. *Ecology* **53,** 22–45 (1972).
15. Smith, W. K. Temperatures of desert plants—another perspective on adaptability of leaf size. *Science* **201,** 614–616 (1978).
16. Helliker, B. R. & Ehleringer, J. R. Establishing a grassland signature in veins: ^{18}O in the leaf water of C_3 and C_4 grasses. *Proc. Natl Acad. Sci. USA* **97,** 7894–7898 (2000).
17. Wang, X.-F. & Yakir, D. Temporal and spatial variations in the oxygen-18 content of leaf water in different plant species. *Plant Cell Environ.* **18,** 1377–1385 (1995).
18. Buhay, W. M., Edwards, T. W. D. & Aravena, R. Evaluating kinetic fractionation factors used for ecologic and paleoclimatic reconstructions from oxygen and hydrogen isotope ratios in plant water and cellulose. *Geochim. Cosmochim. Acta* **60,** 2209–2218 (1996).
19. Barbour, M. M. & Farquhar, G. D. Relative humidity- and ABA-induced variation in carbon and oxygen isotope ratios of cotton leaves. *Plant Cell Environ.* **23,** 473–485 (2000).
20. Richter, S. L., Johnson, A. H., Dranoff, M. M. & Taylor, K. D. Continental-scale patterns in modern wood cellulose $\delta^{18}O$: implications for interpreting paleo-wood cellulose $\delta^{18}O$. *Geochim. Cosmochim. Acta*. (in the press).
21. Sternberg, L. D. L. *et al.* Oxygen isotope ratios of cellulose-derived phenylglucosazone: An improved paleoclimate indicator of environmental water and relative humidity. *Geochim. Cosmochim. Acta* **71,** 2463–2473 (2007).
22. Smith, W. K. & Carter, G. A. Shoot structural effects on needle temperatures and photosynthesis in conifers. *Am. J. Bot.* **75,** 496–500 (1988).
23. Leuzinger, S. & Korner, C. Tree species diversity affects canopy leaf temperatures in a mature temperate forest. *Agric. For. Meteorol.* **146,** 29–37 (2007).
24. Michaletz, S. T. & Johnson, E. A. Foliage influences forced convection heat transfer in conifer branches and buds. *New Phytol.* **170,** 87–98 (2006).
25. Long, S. P. & Woodward, F. I. (eds) *Plants and Temperature* (Society for Experimental Biology, Cambridge, 1988).
26. Berry, J. & Björkman, O. Photosynthetic response and adaptation to temperature in higher plants. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **31,** 491–543 (1980).
27. Way, D. A. & Sage, R. F. Elevated growth temperatures reduce the carbon gain of black spruce [*Picea mariana* (Mill.) B.S.P.]. *Glob. Change Biol.* **14,** 624–636 (2008).
28. Jahren, A. H. & Sternberg, L. S. L. Humidity estimate for the middle Eocene Arctic rain forest. *Geology* **31,** 463–466 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. H. Johnson for discussion of the results; D. Vann and M. Dranoff for help with analysis; and B. Casper, P. Petraitis and D. Brisson for comments on the manuscript. This work was supported by a start-up grant from the University of Pennsylvania and a grant from the A.W. Mellon Foundation.

Author Contributions S.L.R. developed the framework for the sampling scheme and analysed the tree-ring cores. B.R.H. developed the framework for the modelling analysis and wrote the majority of the paper. Both authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to B.R.H (helliker@sas.upenn.edu).

METHODS

Tree species and isotope analysis. Tree-ring wood was obtained from 39 species at 25 sites (see Supplementary Table 1) in eastern North America and the Caribbean. The samples and sites were pulled from a larger meta-analysis[20] of tree-cellulose $\delta^{18}O$ based on nearly equal distribution of summer and winter precipitation to satisfy our assumption that, over the lifetime of a tree, the precipitation-amount-weighted mean of $\delta^{18}O$ was equal to water available for plant uptake ($\delta^{18}O_s$). All sample sites were chosen by the availability of class A climate data from nearby weather stations at similar altitudes. The cores were ground, homogenized and α-cellulose was extracted from a 0.2–0.3-g subsample with a 90:10 mixture of acetic acid (80%; v/v) and nitric acid (69%; v/v) at 120 °C for 2 h. Samples were then washed in ethanol and subsequent ethanol–acetone washes[29]. Additionally, the samples were rinsed with 10% and then 17% NaOH to remove hemicelluloses[27]. Samples were weighed into silver capsules and pyrolysed at 1,100 °C in a Costech Elemental Analyser. The CO gas evolved flowed online into a Thermo-Finnigan Delta Plus isotope ratio mass spectrometer. All samples were run in triplicate. Daily precision of the instrument ranged from 0.06 to 0.35‰, and the standard deviation for reference standards was 0.23‰ ($n = 83$). Isotope ratios (δ) were expressed relative to Vienna Standard Mean Ocean Water (VSMOW) by

$$\delta^{18}O = \left(\frac{R_{sample}}{R_{standard}} - 1\right) \times 1{,}000 \quad (1)$$

where R is the molar isotope ratio of $^{18}O/^{16}O$ and $R_{standard}$ (VSMOW) = 0.0020052.

Isotope discrimination above local precipitation water (the assumed tree source water) was expressed as

$$\Delta = \left(\frac{R_{sample}}{R_{source}} - 1\right) \times 1{,}000 \quad (2)$$

(approximated by $\delta^{18}O_{sample} - \delta^{18}O_{source}$), where R is the molar isotope ratio of either the cellulose or tree source water.

Theory. A previous cellulose $\delta^{18}O$ model[19] was used to predict $\Delta^{18}O_c$ (see Fig. 1):

$$\Delta^{18}O_c = \Delta^{18}O_{lw}(1 - p_{ex}p_x) + \varepsilon_c \quad (3)$$

where $\Delta^{18}O_{lw}$ is the $\delta^{18}O$ of leaf water (relative to plant source water) in which the sucrose substrates for cellulose are synthesized, ε_c is the equilibrium fractionation factor between organically bound oxygen and synthesis water, p_{ex} is the number of organically bound oxygen atoms in leaf-formed sucrose that exchange with xylem water on cellulose formation during tree-ring synthesis (range 0 to 1) and p_x is the proportional deviation of the isotope ratio of xylem water from plant source water (range 0 to 1).

The isotope ratio of water in which the sucrose substrates for cellulose are synthesized ($\Delta^{18}O_{lw}$) is a balance of enriched water at the evaporative site ($\Delta^{18}O_{es}$) and unenriched vein water in the leaf. This balance can be described by a Péclet effect ($\wp = LE/CD$) that accounts for the opposing fluxes of evaporative, convective flux through the leaf (E) across a given path length (L) as opposed to the diffusion of water away from the evaporative sites (CD, where C is the molar density of water and D is the diffusivity of $H_2{}^{18}O$ in water):

$$\Delta^{18}O_{lw} = [\Delta^{18}O_{es}(1 - e^{-\wp})]/\wp \quad (4)$$

The description of ^{18}O enrichment in leaf water at the evaporative site within a leaf can be described by

$$\Delta^{18}O_{es} = \varepsilon^* + \varepsilon^k\left(\Delta^{18}O_v - \varepsilon^k\right)\frac{e_a}{e_i} \quad (5)$$

(refs 30, 31), where $\Delta^{18}O_v$ is the atmospheric water vapour $\delta^{18}O$ relative to source water and ε^k and ε^* are the kinetic and temperature-dependent equilibrium fractionation factors for water (vapour) diffusion and evaporation. The evaporative gradient of water loss from the leaf to the atmosphere is represented by e_a/e_i, which is the ambient vapour pressure divided by the saturation vapour pressure at leaf temperature.

To solve for tree-canopy leaf temperature (T_L) from the observed cellulose $\Delta^{18}O_c$, equations (4) and (5) were inserted into equation (3), which was rearranged to solve for e_i:

$$e_i = \frac{\left(\Delta^{18}O_v - \varepsilon^k\right)e_a}{\left(\frac{\left(\Delta^{18}O_c - \varepsilon_c\right)\wp}{(1 - p_{ex}p_x)(1 - e^{-\wp})}\right) - \varepsilon^* - \varepsilon^k} \quad (6)$$

T_L was obtained from e_i by rearranging a standard vapour–pressure–temperature relationship[32]:

$$T_L = \frac{240.97\left(\ln\frac{e_i}{0.61365}\right)}{17.502 - \left(\ln\frac{e_i}{0.61365}\right)} \quad (7)$$

Model parameterization. To develop predictions for $\Delta^{18}O_c$ (open squares in Fig. 1) we used the average of published observations for the parameters in equations (1)–(3) and used the range of these observations to develop the largest possible range of predictions (error bars in Fig. 1; see also Supplementary Table 1). The oxygen isotope ratio of observed and predicted cellulose ($\Delta^{18}O_c$) were presented relative to tree source water to highlight tree-specific isotopic enrichment above a common source water at a given site. Tree source water was assumed to be equal to the precipitation-weighted model outputs in ref. 7. These model outputs are based on observations from the Global Network of Isotopes in Precipitation (http://isohis.iaea.org) and have been shown to correspond well to the observed $\delta^{18}O$ in tree source water over a variety of sites[21]. To develop a range for source water inputs, we assumed a ±15% error in this source water value. We used $p_{ex}p_x = 0.4$ and $\varepsilon_C = 27$‰ (refs 2, 12, 33) and varied $p_{ex}p_x$ from 0.38 to 0.42, the range observed for trees in both tightly controlled greenhouse conditions and natural field conditions. The average atmospheric water vapour $\delta^{18}O$ relative to source water ($\Delta^{18}O_v$) was determined by assuming equilibrium with amount-weighted precipitation inputs ($\delta^{18}O_s$) at the mean growing-season temperature. As temperature varies seasonally, the equilibrium fractionation between precipitation water and water vapour will change, therefore the value of $\Delta^{18}O_v$ important to leaf water enrichment should correspond to growing-season temperatures. To determine the Péclet number ($\wp$) for each species we had to determine a value for transpiration E and the effective path length for diffusion L. We solved for E using $E = g\,(e_i - e_a)/P$, where g is stomatal conductance to water vapour and P is atmospheric pressure. The value of g was obtained from global estimates by biome and within a biome by gymnosperm versus angiosperm[34], weather-station relative humidity was used to determine e_i and e_a, and P was determined by sample site altitude. Observed values of L have ranged from 0.004 to 0.240 mm. However, studies that have determined L from sucrose $\delta^{18}O$, and not just the offset of observed bulk leaf water from predictions of equation (3), have found L to be consistently smaller than the high end of this range. We used a value for L of 0.015 m and a range of 0.004–0.05 m. For our predictions in Fig. 1, we assumed that leaf temperature equalled ambient temperature and e_a/e_i could be obtained from observed relative humidity ($= 100 \times e_a/e_i$). We used growing season values of monthly mean temperature and relative humidity from the weather-station data, weighted by monthly net primary productivity, to drive the models. For a particular site the growing season length was determined by estimates of growing degree days[35]. It should be noted that predicted cellulose and leaf temperature values vary little if only the monthly means of July and August are used for growing-season temperature at each site. In summary, the error bars for the predictions in Fig. 1 are the upper-end and lower-end predictions using the following ranges: $p_{ex}p_x = 0.38$–0.42 and $L = 0.004$–0.05 mm, given a ±15% error in $\delta^{18}O$ in plant source water.

We first solved for e_i (and subsequently T_L) by using growing-season air temperature as the initial determinant of ε^*—the only temperature-controlled variable on the right-hand-side of equation (6). The initial calculation of T_L was then used to determine ε^* for the second iterative determination of e_i and T_L. Four iterations were performed to arrive at the final T_L. On average, T_L differed by 0.3 °C from the final value after the second iteration and by 0.03 °C after the third iteration. It should be noted that changes in e_i would force changes in plant transpiration (E) because we used a constant canopy conductance to water vapour. Such a change in E could force a change in the Péclet number, $\wp$. However, we maintained a constant $\wp$ to solve for canopy leaf temperature because updating E and $\wp$ at every iterative step for every species and site was computationally difficult. We did, however, perform this exercise with a few select species and found no significant difference between solved values of T_L.

29. Brendel, O., Iannetta, P. P. M. & Sterwart, D. A rapid and simple method to isolate pure α-cellulose. *Phytochem. Anal.* **11**, 7–10 (2000).
30. Craig, H. & Gordon, L. I. in *Stable Isotopes in Oceanographic Studies and Paleotemperatures* (ed. Tongiorigi, E.) 9–130 (Consiglio Nazionale Delle Ricerche Laboratorio di Geologia Nucleare, Pisa, 1965).
31. Farquhar, G. D. & Lloyd, J. in *Stable Isotopes and Plant Carbon/Water Relations* (eds Ehleringer, J. R., Hall, A. E. & Farquhar, G. D.) 47–70 (Academic, San Diego, CA, 1993).
32. Buck, A. L. New equations for computing vapor pressure and enhancement factor. *J. Appl. Meteorol.* **20**, 1527–1532 (1981).
33. Sternberg, L. S. L. in *Stable Isotopes in Ecological Research* (eds Rundel, P. W., Ehleringer, J. R. & Nagy, K. A.) 124–141 (Springer, New York, 1989).
34. Schulze, E.-D., Kelliher, F. M., Korner, C., Lloyd, J. & Leuning, R. Relationships among maximum stomatal conductance, ecosystem surface conductance, carbon assimilation rate, and plant nitrogen nutrition: A global ecology scaling exercise. *Annu. Rev. Ecol. Syst.* **25**, 629–660 (1994).
35. Prentice, I. C. *et al.* A global biome model based on plant physiology and dominance, soil properties and climate. *J. Biogeogr.* **19**, 117–134 (1992).

LETTERS

Ecosystem energetic implications of parasite and free-living biomass in three estuaries

Armand M. Kuris[1]*, Ryan F. Hechinger[1]*, Jenny C. Shaw[1], Kathleen L. Whitney[1], Leopoldina Aguirre-Macedo[2], Charlie A. Boch[1], Andrew P. Dobson[3], Eleca J. Dunham[4], Brian L. Fredensborg[5], Todd C. Huspeni[6], Julio Lorda[1], Luzviminda Mababa[1], Frank T. Mancini[7], Adrienne B. Mora[8], Maria Pickering[9], Nadia L. Talhouk[1], Mark E. Torchin[10] & Kevin D. Lafferty[11]

Parasites can have strong impacts but are thought to contribute little biomass to ecosystems[1–3]. We quantified the biomass of free-living and parasitic species in three estuaries on the Pacific coast of California and Baja California. Here we show that parasites have substantial biomass in these ecosystems. We found that parasite biomass exceeded that of top predators. The biomass of trematodes was particularly high, being comparable to that of the abundant birds, fishes, burrowing shrimps and polychaetes. Trophically transmitted parasites and parasitic castrators subsumed more biomass than did other parasitic functional groups. The extended phenotype biomass controlled by parasitic castrators sometimes exceeded that of their uninfected hosts. The annual production of free-swimming trematode transmission stages was greater than the combined biomass of all quantified parasites and was also greater than bird biomass. This biomass and productivity of parasites implies a profound role for infectious processes in these estuaries.

Standing stock biomass and biomass production are traditional measures of the energetics of ecosystems (see, for example, refs 4–6). Infectious agents are perceived to contribute negligible biomass to ecosystems[1–3]. If so, it may be appropriate to set them aside from investigations of energetics, ecosystems or food webs. However, some parasites markedly influence host individuals (notably humans), wildlife populations and sometimes host communities. These effects imply a general role for infectious processes in the dynamics of ecosystems. Here we quantify the biomass of free-living organisms and their parasites in three estuaries.

Over the course of five years we performed an extensive quantification of the free-living and infectious biomass in three estuaries in Baja California (Bahia Falsa in Bahia San Quintín (BSQ) and Estero de Punta Banda (EPB)) and California (Carpinteria Salt Marsh (CSM)). Cumulatively, the study included 199 species of free-living animals, 15 species of free-living vascular plants and 138 species (including 1 plant species) of infectious agents (see Table 1). Unless specifically mentioned, biomass refers to wet weight, including hard parts.

Here we consider the biomass of free-living and parasitic species grouped by taxonomic categories and, for parasites, by life-history strategy[7,8]. We also determined the proportion of the mass in each host category that was parasite tissue. Additionally—because several parasites in our study were parasitic castrators, usurping the phenotype of their hosts—we noted the biomass in each estuary of castrated hosts (parasite extended phenotypes[9]). Trematode castrators in snail intermediate hosts contributed the most substantial parasitic standing crop biomass in these estuaries, so we further estimated the rates of annual productivity for this infectious component of the system (asexual production of cercariae). To illustrate more sharply the importance of parasite biomass, we compare it directly with the biomass of free-living groups, particularly with that of the bird

Table 1 | Summary of free-living groups and animal parasite functional groups in this study, and number of hosts dissected

Free-living group	No. of species	No. of individuals dissected	No. of parasite species				
			Macroparasites	Trophically transmitted	Castrators	Pathogens	Sum
Miscellaneous phyla	10	55	–	1	–	–	1
Small arthropods	33	258	–	2	–	–	2
Polychaetes	38	533	1	7	–	*2*	10
Bivalves	15	267	2	15	1	*1*	19
Snails	11	14,158	–	9	24	*1*	34
Burrowing shrimps	2	87	1	7	2	–	10
Crabs	3	949	1	19	2	*6*	28
Fishes	17	965	6	19	–	1	26
Birds	70	162	30	–	–	–	30
Total host–parasite combinations	–	–	41	79	29	11	160
Total species, life stages or individuals	199	17,434	40	72	29	9	150

Totals for numbers of parasite species may be less than the sum of the rows because some parasite species use more than one host group. Italic numbers indicate species for which we did not quantify biomass.

[1]Department of Ecology, Evolution and Marine Biology and Marine Science Institute, University of California, Santa Barbara, California 93106, USA. [2]Centro de Investigación y Estudios Avanzados del IPN, C.P. 97310, Mérida, Mexico. [3]Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08544-1003, USA. [4]Center for Infectious Disease Dynamics, Department of Biology, Pennsylvania State University, University Park, Pennsylvania 16802, USA. [5]Department of Biology, University of Texas Pan-American, Edinburg, Texas 78539, USA. [6]Department of Biology, University of Wisconsin–Stevens Point, Stevens Point, Wisconsin 54481, USA. [7]Pacific Islands Fisheries Research Center, National Marine Fisheries Service, Honolulu, Hawaii 96822, USA. [8]Department of Biology, University of California, Riverside, California 92521, USA. [9]Ecology and Evolutionary Biology, University of Connecticut, Storrs, 75 North Eagleville Rd. Unit 3043, Storrs, Connecticut 06269, USA. [10]Smithsonian Tropical Research Institute, Apartado 0843, Ancon, Balboa 03092, Panama, Republic of Panama. [11]Western Ecological Research Center, US Geological Survey, Marine Science Institute, University of California, Santa Barbara, California 93106, USA.
*These authors contributed equally to this work.

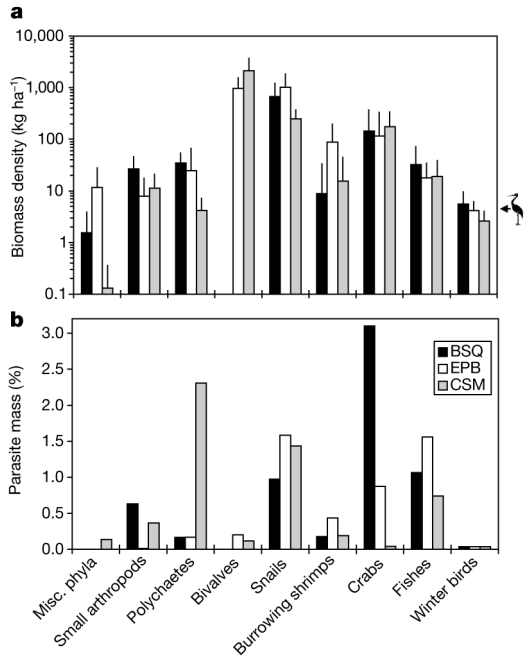


Figure 1 | Biomass of animals and proportional contribution of parasites in three estuaries. **a**, Ecosystem-level biomass density of free-living animal groups. **b**, Parasite tissue as a percentage of total biomasses. The arrow at the bird icon in **a** marks the mean biomass of winter birds (4.1 kg ha^{-1}) across the three estuaries. Error bars in **a** indicate upper 95% confidence limit. The Supplementary Information contains the standard errors and degrees of freedom for the stratified means, and confidence limits for this and all other figures.

assemblage (an obvious and important component of the estuarine ecosystem that includes most of the top predators[10]).

Vascular plants composed the greatest fraction of the biomass in all three estuaries: a mean of 136,166 ± 33,848 (95% confidence limits) kg ha^{-1} at BSQ, 61,754 ± 14,512 kg ha^{-1} at EPB, and 169,035 ± 26,606 kg ha^{-1} at CSM. At CSM, the parasitic dodder, *Cuscuta salina*, infecting leaves and stems, was 0.27% of the plant biomass. Dodder was less common at EPB and scarce at BSQ. We recognized 199 species of free-living animals and 150 species (or life stages) of metazoan parasites (Table 1).

Faunal composition was similar across these estuaries. As regards the species that contribute the top 95% of all biomass, 28% of free-living and 71% of parasite species were common to all three estuaries, and 67% of free-living species and 74% of parasite species were common to at least two estuaries. The biomasses of all free-living animals (including their infectious agents) were 925 kg ha^{-1} at BSQ, 2,240 kg ha^{-1} at EPB, and 2,594 kg ha^{-1} at CSM. In the three estuaries, parasites composed 1.2%, 0.9% and 0.2% of the total animal biomass, respectively. Additionally, parasite biomasses were 6.3%, 13.2% and 3.2% of the combined biomass of their free-living trophic counterparts—that is, the main free-living groups that also feed on multiple trophic levels, namely crabs, fishes, miscellaneous phyla and birds.

For visual presentation, we combined free-living species into broad taxonomic categories (Fig. 1a and Table 1). Our estimates for free-living biomass compare with those from other estuaries (see, for example, refs 11–13). The most substantial contributors to animal biomass were the snails, bivalves and crabs. Across estuaries, the biomasses of the broad categories were generally consistent, the striking exception being the lack of bivalves at BSQ. Other biomass differences between estuaries were driven to a large extent by differences in relative habitat areas (for example marsh habitat, which was relatively extensive at CSM, supported fewer fishes and invertebrates). When all parasites were combined within the free-living groups, the total mass of parasites was generally less than 2% of the biomass of their host categories (Fig. 1b). However, the percentage of parasite biomass varied between estuaries and sometimes reached more than 3% of the mass of their free-living host groups.

The average parasite group had a biomass three orders of magnitude lower than that of the average free-living group (Fig. 2a). Certain parasitic groups dominated the parasite biomass, reaching levels similar to those of common free-living groups. For instance, the biomass of trematode worms was comparable to that of the fishes, burrowing shrimps, polychaetes or small arthropods. In all estuaries, trematode biomass exceeded bird biomass by threefold to ninefold. The epicaridean isopods were the second biggest biomass component of the parasite groups (along with tapeworms at BSQ and EPB). As with the free-living groups, biomass estimates for parasite groups were similar for all estuaries, with exceptions being the small contribution of cestodes and parasitic copepods at CSM.

Parasitic castrators and trophically transmitted parasite stages dominated parasite biomass, attaining 1–10 kg ha^{-1} (Fig. 2b). This mass density was comparable to—or exceeded—that of the vertebrate groups in these estuaries. Macroparasites contributed much less to estuary biomass. This was partly due to the relatively low biomass of their principal hosts (birds and fishes). The total biomasses of the functional groups of parasites were also similar across estuaries.

A host infected with a parasitic castrator has the effective genotype of the parasite[14]. Hence, the entire mass of each castrated host constitutes the extended phenotype[9] of its parasitic castrator. For a host group, the biomass of parasitically castrated hosts approached and sometimes exceeded the biomass of their uninfected hosts (Fig. 3).

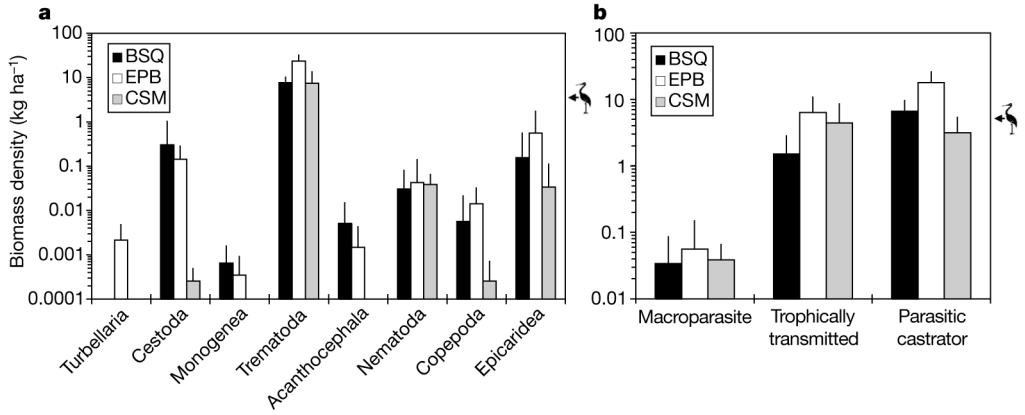


Figure 2 | Ecosystem-level biomass density of animal parasites in three estuaries. **a**, Parasites grouped by major taxon. **b**, Parasites grouped by functional group. The reference arrow at the bird icon marks the mean winter bird mass density across the three estuaries (4.1 kg ha^{-1}). Error bars indicate upper 95% confidence limit.

For example, across the three estuaries, parasitically castrated *Cerithidea californica* commandeered 37–130% of the soft-tissue biomass compared to the uninfected snail populations (Fig. 4). Thus, parasites effectively controlled much of the host biomass of some free-living groups. This probably applies to the many other marine and aquatic systems in which hosts for parasitic castrators (for example crabs, shrimps and snails) are common.

The snail *C. californica* and its larval trematode parasitic castrators were considerable components of animal biomass. *C. californica* had the greatest biomass of any invertebrate in the two southern estuaries (569 kg ha^{-1} at BSQ, 854 kg ha^{-1} at EPB) and ranked eighth among the invertebrates at CSM (144 kg ha^{-1}). The larval parthenitae of 18 recognized trematode species parasitically castrated many of these snails, including almost all of the largest individuals. The trematodes average 22% of the total soft-tissue weight of individual infected snails[15]. In total, the trematode biomass in *C. californica* matched or exceeded the high winter biomass of birds and substantially exceeded their summer biomass (Fig. 4).

We quantified the combined cercarial production of the 18 trematode species infecting *C. californica* snails. Because their snail hosts were large and abundant, these cercariae comprised a substantial component of parasite productivity. Cercariae are released from snails in a daily pulse[16] and have ephemeral life spans of about 24 h. The annual cercarial biomass produced by all *C. californica* trematodes could therefore be compared with the standing crop biomass of other (long-lived) animals. Annual production of cercariae was about threefold that of trematode parthenitae standing-stock biomass and threefold to tenfold that of winter bird biomass (Fig. 4). Further the annual production of cercariae exceeded 1.3–2.2-fold the standing stock of all parasites combined. Reproductive effort—the biomass of offspring (cercariae) produced in a year divided by the biomass of parents (infected snail soft-tissue mass)—was 0.53–0.86. This reproductive effort lies outside the range of values (0.065–0.29) reported for 13 iteroparous marine mollusc species[17,18]. Both parthenitae in *C. californica* and cercariae produced by trematodes infecting *C. californica* had greater densities in the two southern estuaries, primarily as a result of the abundance of *C. californica* throughout the vegetated marsh at BSQ and EPB, whereas at CSM snails were rare in this extensive habitat (50–53% of all habitat area at BSQ and EPB, and 77% of that at CSM).

Our conservative estimates (see Methods) indicate that parasite biomass is comparable to that of several major groups of free-living animals and greater than that of the principal top predators in these estuaries. Parasite biomass was not equally distributed among host or parasite groups; the parasitic castrator functional group comprised most of the parasitic biomass. Consideration of the influence of their extended phenotypes indicates a large ecological role for such parasites. Further, parasite biomass relative to the free-living biomass was up to 6–12-fold the 0.1–0.2% 'best guess' used for an ecosystem model of coral reefs that predicted a significant increase in trophic efficiency when parasites were included in the model[19].

Large standing-stock biomass is not the only indication of energetic importance to ecosystems: productivity is also fundamental[4,5,20]. Parasites efficiently convert food to growth and reproduction, perhaps because they are released from the homeostatic, food gathering and mobility tasks conducted by their hosts[21]. Thus, parasites—such as larval trematodes in snails—may generally have substantial biomass (like many macroorganisms) and high productivity (like microbial organisms).

Additionally, parasites drain host energy beyond that which they consume. Resistance to parasites can be energetically costly (as a result of physiological and behavioural traits to detect, prevent and respond to infection)[22]. In particular, immune systems require substantial standing investment and incur inductive energetic costs[23], and added to that are costs of repairing tissue damaged or consumed by parasites. If parasites have relatively high productivity compared with free-living consumers, and non-consumptive effects on their resources, their effects at the ecosystem level could be disproportionately greater than suggested by their biomass.

This investigation of the biomass of parasites at the ecosystem level fits with emerging interest in the role of parasites in food webs. Parasites can significantly affect food-web topology (for example, increasing chain length and connectance) and are commonly consumed[24,25]. Further, by modifying the behaviour of intermediate hosts, parasites can selectively strengthen links between predator and prey[26]. A quantification of biomass allows the assignment of mass to these potentially important parasitic nodes and therefore represents a step towards fully dynamic food-web models that incorporate infectious processes.

The substantial biomass and productivity attributed to parasites in these estuaries calls for the full integration of parasite ecology into the general body of ecological theory. Food-web analyses and ecosystem

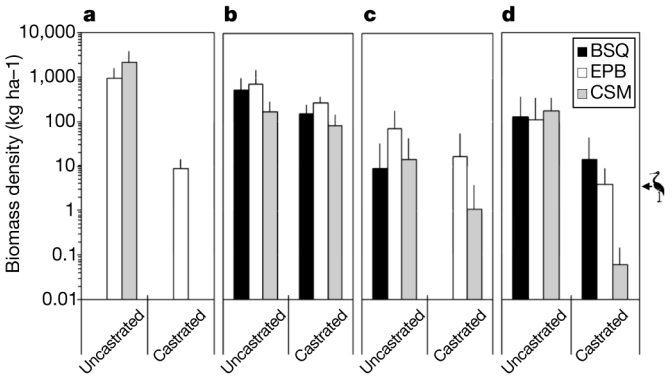


Figure 3 | Ecosystem-level biomass density of parasitically castrated (extended phenotypes) and uninfected phenotypes of hosts supporting parasitic castrators in three estuaries. **a**, Bivalves. **b**, Snails. **c**, Burrowing shrimps. **d**, Crabs. The reference arrow at the bird icon marks the mean winter bird mass density across the three estuaries (4.1 kg ha^{-1}). Error bars indicate upper 95% confidence limit.

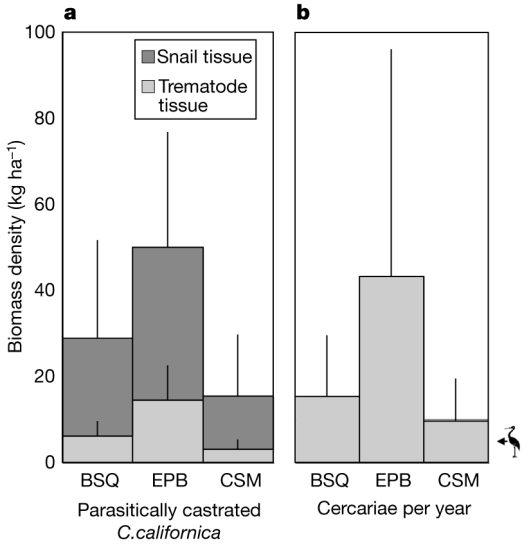


Figure 4 | Standing crop biomass and cercarial productivity of trematodes in *Cerithidea californica* snails. **a**, Ecosystem-level biomass density of host and parasite tissues of parasitically castrated *C. californica*. **b**, Biomass density of the free-swimming stages (cercariae) produced annually by infected snails. Uninfected *C. californica* tissue biomass was 78.8 ± 73.4 (95% confidence limits) kg ha^{-1} at BSQ, 110.5 ± 119.4 kg ha^{-1} at EPB, and 11.8 ± 9.0 kg ha^{-1} at CSM. For clarity we do not include the snail shell mass, which is about 80% of the total mass. The reference arrow at the bird icon marks the mean winter bird-mass density across the three estuaries (4.1 kg ha^{-1}). Summer bird biomass is 0.89 kg ha^{-1} across the three estuaries. Error bars indicate upper 95% confidence limit.

modelling that include parasites[19,24,25,27,28] provide a starting point for this theoretical expansion.

METHODS SUMMARY

We quantified animal and plant wet biomass by sampling 23 random sites in each estuary, stratified over the four major habitats (vegetated marsh, pans, channels, and mudflats and sandflats). At each site we sampled the density and sizes of most free-living organisms more than 1 mm in body size: birds with visual surveys, fishes with nets, benthos with quadrats and cores, and plants with clip quadrats and cores. We estimated free-living animal biomass by applying weight–length curves to the sampled individuals (for birds we used average adult weight).

From each sample site we examined fishes and invertebrates for a wide range of infectious agents, focusing on metazoans. We examined all soft-tissue types in squash preparations. Ethical and pragmatic issues prevented extensive sampling of most bird species for parasites, so we performed a partial estimation of parasite communities of birds by using our own dissections and published information. In general, our methodology probably underestimated the presence of infectious disease (for example, by excluding many pathogens).

We estimated parasite biomass in our samples by multiplying species-specific estimates of individual parasite mass by their abundance[29] in individual hosts. We obtained the masses of most metazoan parasites by directly weighing individuals, or by estimating their mass by multiplying an estimate of their volume by a tissue density of $1.1\ \mathrm{g\,ml^{-1}}$ (ref. 30). To generate estimates for the abundance of parasites in hosts (other than birds), we used statistical models based on data from our dissected hosts.

We estimated the annual productivity of trematode cercariae by multiplying species-specific estimates of individual cercaria mass by species-specific estimates of mean number of cercariae shed daily multiplied by infection density multiplied by 365 days.

Received 3 January; accepted 2 April 2008.

1. Loreau, M., Roy, J. & Tilman, D. in *Parasitism and Ecosystems* (eds Thomas, F., Renaud, F. & Guégan, J.-F.) 13–21 (Oxford Univ. Press, Oxford, 2005).
2. Polis, G. A. & Strong, D. R. Food web complexity and community dynamics. *Am. Nat.* **147,** 813–846 (1996).
3. Poulin, R. The functional importance of parasites in animal communities: many roles at many levels? *Int. J. Parasitol.* **29,** 903–914 (1999).
4. Linderman, R. L. The trophic–dynamic aspect of ecology. *Ecology* **23,** 399–418 (1942).
5. Odum, E. P. Strategy of ecosystem development. *Science* **164,** 262–270 (1969).
6. Yodzis, P. & Innes, S. Body size and consumer–resource dynamics. *Am. Nat.* **139,** 1151–1175 (1992).
7. Kuris, A. M. & Lafferty, K. D. in *Evolutionary Biology of Host–Parasite Relationships: Theory Meets Reality* (eds Poulin, R., Morand, S. & Skorping, A.) 9–26 (Elsevier, Amsterdam, 2000).
8. Lafferty, K. D. & Kuris, A. M. Trophic strategies, animal diversity and body size. *Trends Ecol. Evol.* **17,** 507–513 (2002).
9. Dawkins, R. *The Extended Phenotype: The Long Reach of the Gene* (Oxford Univ. Press, Oxford, 1982).
10. Erwin, R. M. Dependence of waterbirds and shorebirds on shallow-water habitats in the mid-Atlantic coastal region: An ecological profile and management recommendations. *Estuaries* **19,** 213–219 (1996).
11. Spruzen, F. L., Richardson, A. M. M. & Woehler, E. J. Spatial variation of intertidal macroinvertebrates and environmental variables in Robbins Passage wetlands, NW Tasmania. *Hydrobiologia* **598,** 325–342 (2008).
12. Allen, L. G. Seasonal abundance composition and productivity of the littoral fish assemblage in Upper Newport Bay, California. *Fish. Bull.* **80,** 769–790 (1982).
13. Ramer, B. A., Page, G. W. & Yoklavich, M. M. Seasonal abundance habitat use and diet of shorebirds in Elkhorn Slough, California. *Western Birds* **22,** 157–174 (1991).
14. O'Brien, J. & Van Wyk, P. in *Crustacean Issues: Factors in Adult Growth* (ed. Wenner, A.) 191–218 (Balkema, Rotterdam, 1985).
15. Hechinger, R. F. *et al.* How large is the hand in the puppet? Ecological and evolutionary effects on body mass of 15 trematode parasitic castrators in their snail host. *Evol. Ecol.*. doi:10.1007/s10682-008-9262-4 (in the press).
16. Fingerut, J. T., Zimmer, C. A. & Zimmer, R. K. Patterns and processes of larval emergence in an estuarine parasite system. *Biol. Bull.* **205,** 110–120 (2003).
17. Browne, R. A. & Russell-Hunter, W. D. Reproductive effort in molluscs. *Oecologia* **37,** 23–27 (1978).
18. Hughes, R. N. & Roberts, D. J. Reproductive effort of winkles (*Littorina* spp.) with contrasted methods of reproduction. *Oecologia* **47,** 130–136 (1980).
19. Arias-Gonzalez, J. E. & Morand, S. Trophic functioning with parasites: a new insight for ecosystem analysis. *Mar. Ecol. Prog. Ser.* **320,** 43–53 (2006).
20. McLusky, D. S. *The Estuarine Ecosystem* 2nd edn (Blackie, Glasgow, 1989).
21. Calow, P. Pattern and paradox in parasite reproduction. *Parasitology* **86,** 197–207 (1983).
22. Rigby, M. C., Hechinger, R. F. & Stevens, L. Why should parasite resistance be costly? *Trends Parasitol.* **18,** 116–120 (2002).
23. Lochmiller, R. L. & Deerenberg, C. Trade-offs in evolutionary immunology: just what is the cost of immunity? *Oikos* **88,** 87–98 (2000).
24. Lafferty, K. D., Dobson, A. P. & Kuris, A. M. Parasites dominate food web links. *Proc. Natl Acad. Sci. USA* **103,** 11211–11216 (2006).
25. Lafferty, K. D. *et al.* in *Disease Ecology: Community Structure and Pathogen Dynamics* (eds Collinge, S. K. & Ray, C.) 119–134 (Oxford Univ. Press, Oxford, 2006).
26. Lafferty, K. D. & Morris, A. K. Altered behavior of parasitized killifish increases susceptibility to predation by bird final hosts. *Ecology* **77,** 1390–1397 (1996).
27. Huxham, M. & Raffaelli, D. Parasites and food-web patterns. *J. Anim. Ecol.* **64,** 168–176 (1995).
28. Thompson, R. M., Mouritsen, K. N. & Poulin, R. Importance of parasites and their life cycle characteristics in determining the structure of a large marine food web. *J. Anim. Ecol.* **74,** 77–85 (2005).
29. Bush, A. O., Lafferty, K. D., Lotz, J. M. & Shostak, A. W. Parasitology meets ecology on its own terms: Margolis *et al.* revisited. *J. Parasitol.* **83,** 575–583 (1997).
30. Peters, R. H. *The Ecological Implications of Body Size* (Cambridge Univ. Press, Cambridge, 1983).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank many assistants, in particular I. Jimenez, A. Kaplan, M. Saunders, J. Smith, A. Wood and the research team of L. Ladah. L. Ladah and the Huttinger family provided facilities for fieldwork. Satellite imagery of CSM was provided by K. Clarke. The University of California Natural Reserve System provided access to CSM. The National Science Foundation/National Institutes of Health Ecology of Infectious Diseases Program provided funding.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to A.M.K. (kuris@lifesci.ucsb.edu).

LETTERS

Evidence for the evolutionary nascence of a novel sex determination pathway in honeybees

Martin Hasselmann[1]*, Tanja Gempe[1]*, Morten Schiøtt[1,2], Carlos Gustavo Nunes-Silva[3], Marianne Otte[1] & Martin Beye[1]

Sex determination in honeybees (*Apis mellifera*) is governed by heterozygosity at a single locus harbouring the *complementary sex determiner* (*csd*) gene[1], in contrast to the well-studied sex chromosome system of *Drosophila melanogaster*[2]. Bees heterozygous at *csd* are females, whereas homozygotes and hemizygotes (haploid individuals) are males. Although at least 15 different *csd* alleles are known among natural bee populations[3], the mechanisms linking allelic interactions to switching of the sexual development programme are still obscure. Here we report a new component of the sex-determining pathway in honeybees, encoded 12 kilobases upstream of *csd*. The gene *feminizer* (*fem*) is the ancestrally conserved progenitor gene from which *csd* arose and encodes an SR-type protein, harbouring an Arg/Ser-rich domain. Fem shares the same arrangement of Arg/Ser- and proline-rich-domain with the *Drosophila* principal sex-determining gene *transformer* (*tra*), but lacks conserved motifs except for a 30-amino-acid motif that Fem shares only with Tra of another fly, *Ceratitis capitata*[4]. Like *tra*, the *fem* transcript is alternatively spliced. The male-specific splice variant contains a premature stop codon and yields no functional product, whereas the female-specific splice variant encodes the functional protein. We show that RNA interference (RNAi)-induced knockdowns of the female-specific *fem* splice variant result in male bees, indicating that the *fem* product is required for entire female development. Furthermore, RNAi-induced knockdowns of female allelic *csd* transcripts result in the male-specific *fem* splice variant, suggesting that the *fem* gene implements the switch of developmental pathways controlled by heterozygosity at *csd*. Comparative analysis of *fem* and *csd* coding sequences from five bee species indicates a recent origin of *csd* in the honeybee lineage from the *fem* progenitor and provides evidence for positive selection at *csd* accompanied by purifying selection at *fem*. The *fem* locus in bees uncovers gene duplication and positive selection as evolutionary mechanisms underlying the origin of a novel sex determination pathway.

We have identified a second switch gene, *fem*, in the honeybee that, besides *csd*, also localizes to the sex determination locus (SDL; Fig. 1a). The SDL defines the genomic region that is always heterozygous in females[5] and thus possibly harbours extra genes involved in sex determination. We isolated a new part of the SDL (26 kilobases (kb)) by assembling sequences of a previously analysed region (21 kb)[1], fragments from shotgun cloning, contigs from the honeybee genome project[6] and site-specific amplicons. In our previous study[1] we failed to isolate more parts of the SDL because SDL sequences are AT-rich and are under-represented in our various cloning and shotgun sequencing strategies[1,6]. We now report three extra genes within the SDL. We tested all five genes located within the SDL (Fig. 1a) for sex-determining function by RNAi knockdown experiments. Only *csd* and the new *fem* gene, located 12 kb upstream of *csd*, have sex determination function (Fig. 1b). RNAi-induced knockdowns of *fem* in females result in a developmental switch to entire male head differentiation (Fig. 1b), whereas knockdowns in males do not affect head development. Repressing the function of *csd* by RNAi results in apparently the same male-like development in females, but again does not affect head differentiation in males (bottom right panel of Fig. 1b). These findings indicate that *fem* is the second binary switch gene of the sex determination pathway that, when active, regulates the entire developmental programme of females but not that of males. The *fem* gene encodes a protein that has a carboxy-terminal Arg/Ser-rich and Pro-rich domain with a high degree of sequence identity to the Csd protein (>70% identical amino acid residues, Fig. 1c), but no similarity to other proteins in the database by applying BLAST program searches. Fem has an extra Arg/Ser-domain in its amino terminus, but lacks the hypervariable region of Csd (Fig. 1c)[3]. Thus, both genes are evidently paralogues coding for SR-type proteins, which are thought to be involved generally in the regulation of RNA splicing[7].

We characterized sex-specific transcripts of *fem* and identified male and female *fem* transcripts with the same 5′ untranslated region (UTR) sequence, but differences in their downstream exon composition (Fig. 2a). Male-specific *fem* transcripts retain a full exon 3, which contains a stop codon, so that translation terminates prematurely. In females, this part of exon 3 plus exons 4 and 5 are spliced out, leading to a complete open reading frame that translates into a protein of 403 amino acids. Consistent with the allelic mode of *csd* activation[1], we isolated *csd* transcripts that differ in their sequence composition, but not in their combination of exons (Fig. 1c). To test whether *fem* in females is regulated by the activity of the *csd* gene, we repressed *csd* in early embryogenesis by RNAi and studied *fem* transcripts in fourth instar larvae. These females have a predominant transcript of the male composition (Fig. 2b), establishing that splicing of *fem* is regulated in response to the function of *csd*. Next, we examined whether transcriptional levels of *csd* are regulated in response to *fem* function. We repressed *fem* in early embryogenesis and measured *csd* messenger RNA levels at the middle stage of embryogenesis. We observed no differences in *csd* mRNA levels between *fem* knockdown and mock short-interfering-RNA-treated (siRNA-treated) embryos ($P > 0.1$, *t*-test; Fig. 2c). Taken together, these results indicate that the binary switch of the sex determination pathway is implemented by alternative splicing of the *fem* transcript in response to heterozygosity at *csd*.

We compared our findings with the *D. melanogaster* pathway. The genes *fem* and *tra* seem to have equivalent functions in sex determination, belong to the same family of SR-type proteins, share the

[1]Department of Genetics, Heinrich Heine University Duesseldorf, Universitaetsstrasse 1, 40225 Duesseldorf, Germany. [2]Centre for Social Evolution, Department of Biology, University of Copenhagen, Universitetsparken 15, 2100 Copenhagen, Denmark. [3]Grupo de Pesquisas em Abelhas (GPA), Instituto Nacional de Pesquisas da Amazônia (INPA) Avenida André Araújo 2936, 69060-001 Manaus, AM, Brazil.
*These authors contributed equally to this work.

same arrangement of regions enriched with arginine and serine (Arg/Ser domain) and proline (Pro-rich region), but harbour no significant identity in sequence motifs (Fig. 1c). Such identify may not be expected given the rapid divergence of this sequence between different dipteran species[4,8]. The *tra* gene of *D. melanogaster* functions as a switch gene, is regulated sex-specifically by alternative splicing and is necessary for the entire development of females[9,10]. It is also part of the sex determination cascade that communicates the upstream X:A (ratio of X chromosomes to autosome sets) and *Sex lethal* (*Sxl*) signal to the downstream gene *doublesex* (*dsx*), which controls the activity of the final target genes necessary for sexual differentiation[11–13]. When we compared Fem with the orthologue of Tra from the mediterranean fly *C. capitata*[4], in addition to the same arrangement of domains (Fig. 1c), we identified a 30-amino-acid motif in which 15 residues are identical (yellow and framed box in Fig. 1c). On the basis of equivalent function and regulation of sex determination, the same arrangement of Arg/Ser- and Pro-enriched domains and the conserved sequence motif, we conclude that *fem* and *tra* have a common evolutionary origin. These homologies establish a common ancestral pathway of sexual regulation at the level of the *tra* gene across insect orders and ~300 million years (Myr) of independent evolution.

We compared the coding sequences of the honeybee paralogues *fem* and *csd* with their orthologues from two related Asian honeybee species (*Apis cerana*, *Apis dorsata*), as well as *fem* sequences from the stingless bee (*Melipona compressipes*) and the bumble bee (*Bombus terrestris*)—both of which are members of the major sister branches of honeybees (corbiculate bees)—and the jewel wasp (*Nasonia vitripennis*), to obtain information on the evolutionary relationship and functional divergence of these genes.

Significantly, the *csd* gene is unique to the honeybee lineage. The gene duplication event can be placed on the phylogeny by comparing the gene family tree with the phylogeny of the bee species (Fig. 3a). The *fem* and *csd* sequences of honeybees form a single clade in the gene tree irrespective of whether we analyse synonymous (Fig. 3a) or amino acid substitutions (Supplementary Fig. 1). We next addressed whether an accelerated substitution rate in the stingless bee sequence offers an alternative explanation to the clustering of honeybee *fem* and *csd* sequences. The stingless bee sequence had a lower relative substitution rate than honeybees (relative rate test, Supplementary Table 1), suggesting that the *fem* and *csd* clade in the gene tree is the consequence of a duplication event within the honeybee lineage. Given some uncertainty in the current phylogenetic relationships among the major sister branches of corbiculate bees, we included

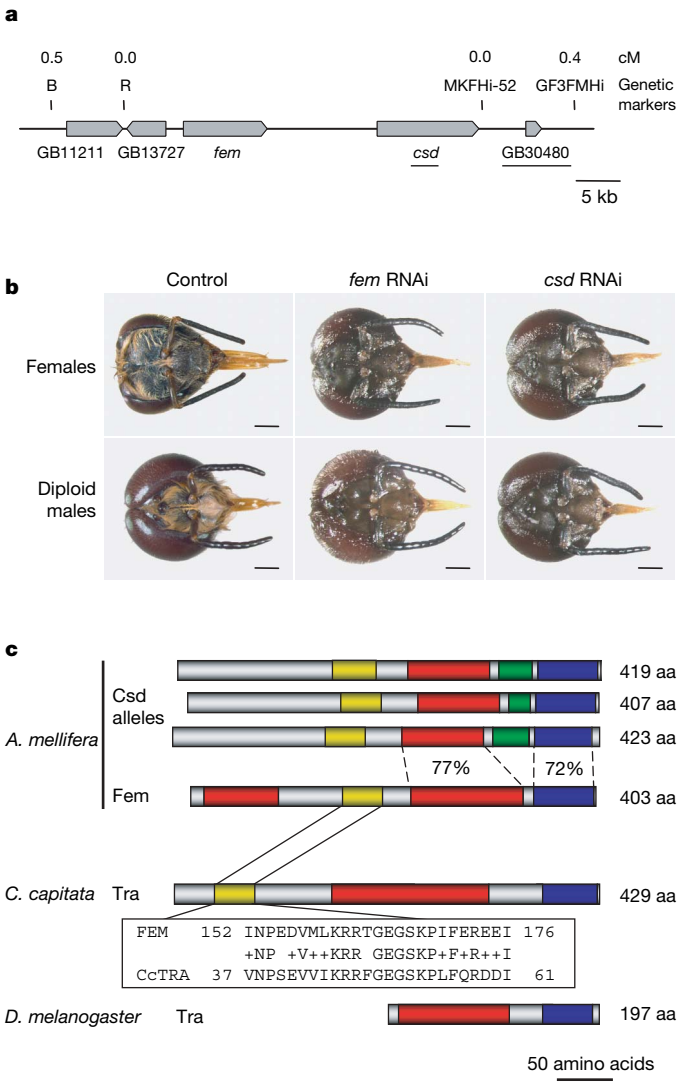


Figure 1 | Sex-determining genes within the SDL genomic region. **a**, Diagram of the identified genes within the SDL genomic region[5]. Genes are orientated 5′ to 3′ according to the direction of the arrows; the names of previously analysed genes[1] are underlined. **b**, Head development of males and females treated with *fem* and *csd* siRNAs in early embryogenesis. Frontal view of female (workers, control on the left, $n = 8$) and diploid male (control on the left, $n = 18$) heads are shown. Seventy-eight per cent ($n = 17$) of females treated with *fem* siRNAs and 75% ($n = 27$) of females treated with *csd* siRNAs develop the entire head structures of males. Diploid males treated with *fem* ($n = 9$) or *csd* ($n = 19$) siRNAs have normally developed male heads. Scale bars, 1 mm. **c**, Domain diagrams of Csd, Fem and Tra proteins. Arg/Ser domains are indicated by red, the hypervariable region by green, and the common Pro-rich region by blue boxes. The sequence motif shared across insect orders (Fem and *C. capitata* Tra (Cc-Tra[4])) and its schematic location (yellow boxes) are shown. The percentage of amino acid identities of Fem and Csd domains are indicated. aa, amino acids.

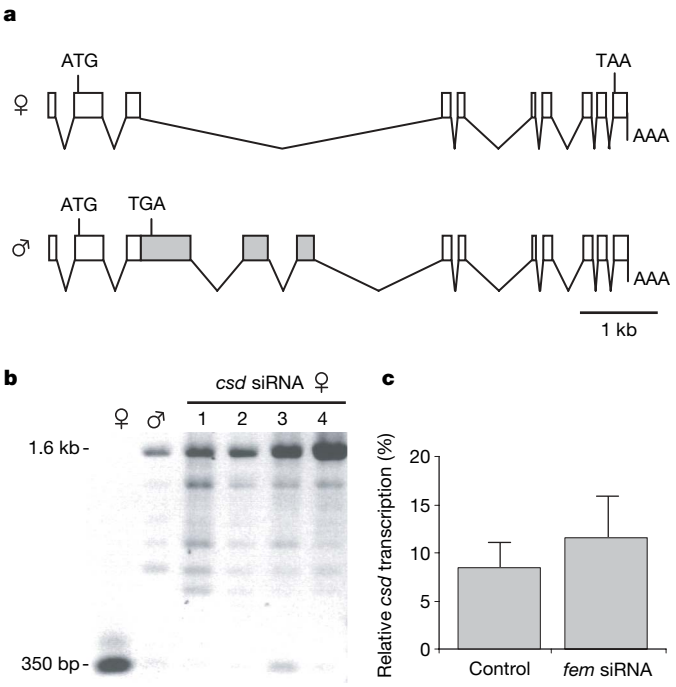


Figure 2 | Structure of *fem* splice variants and the functional relationship of *fem* and *csd* genes. **a**, Female and male splicing diagram of the *fem* gene. Common exons are marked in white, and male-specific exons and exon extension are in grey. Translational start and stop sites as well as the poly(A) addition sites are indicated. **b**, The processing of *fem* transcripts in response to the repression of *csd* function by RNAi. Fragments corresponding to transcripts of female (~350 bp) and male (~1.6 kb) composition were amplified by RT–PCR reactions. **c**, *csd* mRNA amounts in response to repression of *fem* by RNAi in early embryogenesis. Relative transcription amounts of *csd* in *fem* knockdown embryos ($n = 6$) and mock siRNA-treated control embryos ($n = 9$). Error bars represent s.d.

the *N. vitripennis* protein sequence as a further distant out-group (~120 Myr of divergence) in our gene tree analysis (Fig. 3b) and again found a clustering of honeybee Csd and Fem proteins and no evidence for relative differences in substitution rates (relative rate test, Supplementary Table 2). In support of a recent duplication event in the honeybee lineage, we detected only a single copy of this gene family in the bumble bee (*B. terrestris*) and the jewel wasp (*N. vitripennis*). Our Southern blot hybridization of *fem* DNA to *B. terrestris* genomic DNA showed single bands in four different restrictions, supporting the presence of a single genomic locus (Supplementary Fig. 2). By searching the sequenced genome of *N. vitripennis*, we identified only a single gene with homologies to *fem*. The close evolutionary relationship of honeybee *csd* and *fem* genes thus strongly suggests that gene duplication occurred after the split of the stingless bee, bumble bee and honeybee (~70 Myr ago)[14] but before honeybee divergence (~10 Myr ago). Consequently, the *csd* gene is not the universal molecular basis of complementary sex determination in a variety of hymenopteran insects (bees, ants and wasps)[15,16], suggesting that other unknown molecular signals are the primary sex determiners in these species.

Further analysis suggests that the *csd*-based sex determination was shaped by positive selection. An excess of non-synonymous to synonymous substitutions in a branch of a phylogeny indicates that positive darwinian selection has operated in enhancing the fixation of amino acid changes. A significant excess of non-synonymous over synonymous substitutions is observed in the branch immediately after gene duplication, indicating the action of positive selection in the rise of *csd* (Fig. 3c; branch point i–ii, $P = 10^{-4}$, one-tailed Fisher's exact test). This finding is consistent whether we used different honeybee divergence scenarios or included varying numbers of *csd* alleles in the analysis. Six substitutions that are fixed in the *csd* gene are components of a coiled-coil motif which encodes protein-binding properties[17] (Fig. 3d)—a prerequisite for the function of *csd*-based sex determination[18]. In addition to branch i–ii, an excess of non-synonymous over synonymous changes is also found in the branch ii–iii ($P = 0.007$, one-tailed Fisher's exact test) leading to the western honeybee (*A. mellifera*). This latter excess is, however, significantly lower than the former ($P = 0.016$, one-tailed Fisher's exact test) suggesting that stronger positive selection operated during the early formation of the *csd* gene than during lineage-specific divergence. The *fem* gene—the progenitor of *csd*—is under purifying selection, showing an excess of synonymous to non-synonymous substitutions in branches of *fem* ($P < 0.05$, one-tailed Fisher's exact test; Fig. 3c). This low evolutionary divergence of Fem proteins is consistent with our previous result indicating an ancestral sex-determining function for *fem*.

Evidently, *fem* has retained its sex determination function whereas its recent duplicate, *csd*, evolved an allelic mode of SR-type protein activation through positive selection. The *csd*-based sex determination system controls sex-specific splicing of its progenitor transcript, thus implementing the switch of male and female pathways. These findings suggest that gene duplication of an existing major sex

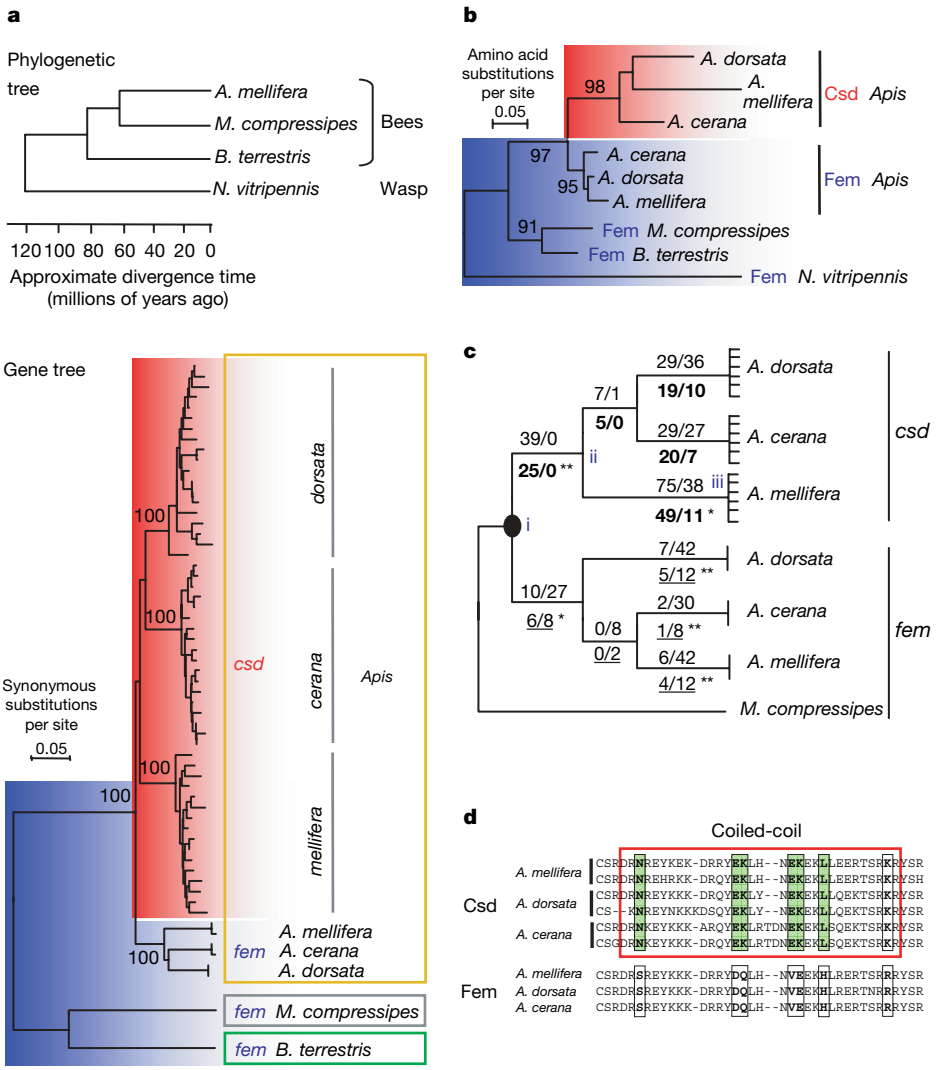


Figure 3 | Nucleotide and amino acid substitutions in the evolution of *csd* and *fem* genes. **a**, Comparison of the phylogenetic[14,30] (top panel) and gene (bottom panel) tree. Branches derived from *csd* and *fem* sequences are indicated by red and blue boxes, respectively. Species sources are marked by coloured frames. Numbers represent bootstrap values >80%. **b**, Gene tree of Fem and Csd protein sequences including the wasp *N. vitripennis*. Numbers denote bootstrap values >80%. **c**, Numbers of non-synonymous (N) and synonymous (S) substitutions along branches of the gene tree. Non-synonymous to synonymous substitution values per branch (a_N/a_S) and per branch and site (b_N/b_S) ×1000 are shown below and on the lines, respectively. Level of significance of Fisher's exact tests for positive (bold numbers) and for purifying (underlined numbers) selection[29]. Single asterisk, $P < 0.05$; double asterisk, $P < 0.01$. i, ii and iii (coloured in blue) define branches between nodes of the *csd* genealogy. **d**, Comparison of fixed amino acid substitutions between Csd and Fem proteins involved in coiled-coil formation. Fixed differences are marked by boxes; the ones in Csd that are components of the coiled-coil motif (red frame) are coloured in green.

determining gene, followed by positive selection in one of the duplicates, favoured the origination of a new upstream signal and, thereby, the creation of a novel sex determination pathway.

The upwards growth of this pathway is consistent with theory[19,20] predicting that new signals are co-opted upstream of a cascade during the course of evolution. Furthermore, it has been proposed that the origin of alternative sex determination signals involve a selective advantage, such as the possibility to modify sex ratios[21] or to improve the quality of signals[21,22]. Our findings provide direct evidence for a role of strong positive selection in the formation of a new sex determination system and are thus consistent with previous suggestions. We propose that in our case the reduction of recombination at the sex determination locus[5] (Fig. 1a) results in a gradual loss of the levels of adaptation of the gene[23,24], which would facilitate the evolution of alternative initial signals of complementary sex determination. This process may thus relate to the evolution of chromosomal systems in which the cessation of recombination[25,26] results in a degradation of genes[27] and the extinction of the sex chromosomes[28]. In either case, our findings provide strong support for the role of positive selection in shaping the growth of a developmental pathway.

METHODS SUMMARY

Queens producing 50% female and 50% diploid males were derived from brother–sister crosses (inbred crosses). Male-producing queens laying exclusively unfertilized, haploid eggs were obtained from non-mated, CO_2-treated queens. Female-producing queens were obtained from queens inseminated by semen of a single male. The sequence reads were assembled using Staden Package software. Potential exons and genes within the SDL were detected by gene prediction programs as described previously[1]. Transcriptionally active genes in embryogenesis were identified by reverse transcription PCR (RT–PCR) of embryonic complementary DNA. The full sequences of transcripts were obtained by 5′ and 3′ RACE experiments. siRNA (50–100 pg per embryo) and double-stranded RNA[1] were injected into embryos at the syncytial stage. Fragments corresponding to sex-specific *fem* transcripts were amplified by RT–PCR reactions from individual fourth instar larvae and resolved by agarose gel electrophoresis. Freshly hatched larvae were reared *in vitro* at 34.5 °C and saturated humidity. Relative transcriptional levels of *csd* were calculated by comparing cycle thresholds to the reference gene, *elongation factor 1-alpha*. Fifty-one *csd* ($n = 15$, *A. mellifera*; $n = 17$, *A. cerana*; $n = 19$, *A. dorsata*) and ten *fem* coding sequences ($n = 4$, *A. mellifera*; $n = 2$, *A. cerana*; $n = 2$, *A. dorsata*; $n = 1$, *M. compressipes*; $n = 1$, *B. terrestris*) were isolated by high-fidelity PCR amplifications of cDNAs derived from RNA preparations of embryos. Gene trees were obtained by the minimum evolution method by applying the Nei–Gojobori distance with Jukes–Cantor correction for nucleotide and Poisson-corrected distances for amino acid differences. The non-synonymous and synonymous substitution values for each branch of the gene tree were obtained by least-squares method. Fisher's exact test for selection was performed on the numbers of changed and unchanged non-synonymous and synonymous sites[29].

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 5 November 2007; accepted 1 May 2008.
Published online 25 June 2008.

1. Beye, M., Hasselmann, M., Fondrk, M. K., Page, R. E. & Omholt, S. W. The gene *csd* is the primary signal for sexual development in the honeybee and encodes an SR-type protein. *Cell* **114,** 419–429 (2003).
2. Cline, T. W. & Meyer, B. J. Vive la différence: males vs females in flies vs worms. *Annu. Rev. Genet.* **30,** 637–702 (1996).
3. Hasselmann, M. & Beye, M. Signatures of selection among sex-determining alleles of the honey bee. *Proc. Natl Acad. Sci. USA* **101,** 4888–4893 (2004).
4. Pane, A., Salvemini, M., Bovi, P. D., Polito, C. & Saccone, G. The *transformer* gene in *Ceratitis capitata* provides a genetic basis for selecting and remembering the sexual fate. *Development* **129,** 3715–3725 (2002).
5. Hasselmann, M. & Beye, M. Pronounced differences of recombination activity at the sex determination locus (SDL) of the honey bee, a locus under strong balancing selection. *Genetics* **174,** 1469–1480 (2006).
6. Honeybee Genome Sequencing Consortium.. Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* **443,** 931–949 (2006).
7. Blencowe, B. J., Bowman, J. A., McCracken, S. & Rosonina, E. SR-related proteins and the processing of messenger RNA precursors. *Biochem. Cell Biol.* **77,** 277–291 (1999).
8. Kulathinal, R. J., Skwarek, L., Morton, R. A. & Singh, R. S. Rapid evolution of the sex-determining gene, *transformer*: structural diversity and rate heterogeneity among sibling species of *Drosophila*. *Mol. Biol. Evol.* **20,** 441–452 (2003).
9. Butler, B., Pirrotta, V., Irminger-Finger, I. & Nothiger, R. The sex-determining gene *tra* of *Drosophila*: molecular cloning and transformation studies. *EMBO J.* **5,** 3607–3613 (1986).
10. Boggs, R. T., Gregor, P., Idriss, S., Belote, J. M. & McKeown, M. Regulation of sexual differentiation in *D. melanogaster* via alternative splicing of RNA from the *transformer* gene. *Cell* **50,** 739–747 (1987).
11. Bell, L. R., Maine, E. M., Schedl, P. & Cline, T. W. *Sex-lethal*, a *Drosophila* sex determination switch gene, exhibits sex-specific RNA splicing and sequence similarity to RNA binding proteins. *Cell* **55,** 1037–1046 (1988).
12. Keyes, L. N., Cline, T. W. & Schedl, P. The primary sex determination signal of *Drosophila* acts at the level of transcription. *Cell* **68,** 933–943 (1992).
13. Hoshijima, K., Inoue, K., Higuchi, I., Sakamoto, H. & Shimura, Y. Control of *doublesex* alternative splicing by *transformer* and *transformer-2* in *Drosophila*. *Science* **252,** 833–836 (1991).
14. Grimaldi, D. & Engel, M. S. *Evolution of the Insects* 454–467 (Cambridge Univ. Press, UK, 2005).
15. Kerr, W. E. Sex determination in bees. XXI. Number of XO-heteroalleles in a natural population of *Melipona compressipes fasciculata* (Apidae). *Insectes Soc.* **34,** 274–279 (1987).
16. Cook, J. M. Sex determination in the Hymenoptera: a review of models and evidence. *Heredity* **71,** 421–435 (1993).
17. Burkhard, P., Stetefeld, J. & Strelkov, S. V. Coiled coils: a highly versatile protein folding motif. *Trends Cell Biol.* **11,** 82–88 (2001).
18. Beye, M. The dice of fate: the *csd* gene and how its allelic composition regulates sexual development in the honey bee, *Apis mellifera*. *Bioessays* **26,** 1131–1139 (2004).
19. Wilkins, A. S. Moving up the hierarchy: A hypothesis on the evolution of a genetic sex determination pathway. *Bioessays* **17,** 71–77 (1995).
20. Nöthiger, R. & Steinemann-Zwicky, M. A single principle for sex determination in insects. *Cold Spring Harb. Symp. Quant. Biol.* **50,** 615–621 (1985).
21. Bull, J. J. *Evolution of Sex Determining Mechanisms* (Benjamin/Cummings Publishing Company, Menlo Park, California, 1983).
22. Pomiankowski, A., Nothiger, R. & Wilkins, A. The evolution of the *Drosophila* sex-determination pathway. *Genetics* **166,** 1761–1773 (2004).
23. Hill, W. G. & Robertson, A. The effect of linkage on limits to artificial selection. *Genet. Res.* **8,** 269–294 (1966).
24. Otto, S. P. & Lenormand, T. Resolving the paradox of sex and recombination. *Nature Rev. Genet.* **3,** 252–261 (2002).
25. Lewis, D. The evolution of sex in flowering plants. *Biol. Rev.* **17,** 46–67 (1942).
26. Charlesworth, B. & Charlesworth, D. A model for the evolution of dioecy and gynodioecy. *Am. Nat.* **112,** 975–997 (1978).
27. Charlesworth, D., Charlesworth, B. & Marais, G. Steps in the evolution of heteromorphic sex chromosomes. *Heredity* **95,** 118–128 (2005).
28. Graves, J. A. Sex chromosome specialization and degeneration in mammals. *Cell* **124,** 901–914 (2006).
29. Nei, M. & Kumar, S. *Molecular Evolution and Phylogenetics* 216–221 (Oxford Univ. Press, New York, 2000).
30. Engel, M. S. Monophyly and extensive extinction of advanced eusocial bees: insights from an unexpected Eocene diversity. *Proc. Natl Acad. Sci. USA* **98,** 1661–1664 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank W. Martin, A. Wilkins, T. Eltz and J. Baines for comments on the manuscript; E.-M. Theilenberg, M. Mueller-Borg and C. Schulte for technical support; D. Titera for providing bee crosses; J. Pflugfelder, N. Koeniger, G. Koeniger, J. Bozic and S. Tingek for collecting honeybee samples; K. Lunau for providing bumble bee samples; and M. Griese for bee-keeping support. This work was supported by grants from the Deutsche Forschungsgemeinschaft DFG.

Author Contributions M.H. performed the evolutionary nucleotide analysis, isolated gene sequences from different species and supervised some experiments; T.G. performed the gene studies; M.S. assembled SDL sequences and identified genes; C.G.N.-S. characterized *M. compressipes* sequences; M.O. analysed domain structures; and M.B. performed the experimental design, supervised the research project and wrote the manuscript.

Author Information The sequences generated in this study are available from GenBank under the accession numbers EU101387 (*GB11211*), EU101388 (*fem*^{F}), EU101389 (*fem*^{M}), EU101390 (*csd*), EU101391 (*GB30480*), EU101392 (*GB13727*), EU139305 (*M. compressipes fem*), EU288185 (*B. terrestris fem*), and EU100885–EU100941 (*Apis fem* and *csd* sequences from populations and different species). SDL assembly and sequence annotation data are available in the Third Party Annotation Section of the DDBJ/EMBL/GenBank databases under the accession number TPA: BK006346. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to M.B. (martin.beye@uni-duesseldorf.de).

METHODS

Bee material. Queens producing 50% female and 50% diploid males were derived from brother–sister crosses (inbred crosses). Male-producing queens were obtained from non-mated, CO_2 treated queens. Female-producing queens were obtained from queens that were inseminated with the semen of a single male. Samples of honeybee species for evolutionary nucleotide analyses were collected in Borneo (Tenom, *A. cerana* and *A. dorsata*), Thailand (Samut Songkram, *A. cerana*; Wanmanaow, *A. dorsata*), Slovenia (Litija, *A. mellifera*) and South Africa (Pretoria, *A. mellifera*). Samples of stingless bees (*Melipona compressipes*) were collected in Brazil (Manaus). Samples of bumble bees (*B. terrestris*) were obtained from the Koppert Company.

Sequence analysis. Sequences of the SDL (AADG05006532, AADG05006533, DQ681226, AY352277 (GenBank accession numbers), and 250636813, 165754571, 566358507, 566314258, 173514040, 566864637, 240540892 and 160908628 (NCBI trace archive numbers)) were assembled using the Staden Package software. Ordering of genomic sequences was verified by cDNA sequences. Potential exons were predicted, and transcriptionally active genes were identified by RT–PCR of embryonic cDNA. The first-strand cDNA was generated by reverse transcription with oligo(dT) primer or random hexamers. RACE experiments were performed to isolate the 5′ and 3′ ends of genes. Sequences were obtained from high-fidelity PCR amplifications of embryonic cDNA and at least three independent clones. We used BLAST programs and a low complexity filter option to compare genes with the database. Domains were analysed by PROSITE database comparisons (http://www.expasy.org/prosite/). The coiled-coil region was predicted by COILS program (http://www.ch.embnet.org/software/COILS_form.html) with a probability of one. Sequences of the coiled-coil motif comparison (Fig. 3d) were derived from GenBank accession numbers: EU100885, EU100893, EU100921, EU100928, EU100904, EU100902, EU100940, EU100938, EU100936. Fem and Csd and Tra proteins (*N. vitripennis* Fem, XM_001604744 and EU780924; *C. capitata* Tra, AF434936; and *D. melanogaster* Tra, P11596) were compared by BLAST2 sequence program (http://www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi).

Functional analysis of genes. RNAi knockdowns were induced in early embryogenesis at the syncytial stage (0–4 h after egg deposition) in haploid and diploid males and females. dsRNAs were generated from cloned cDNAs of genes GB11211, GB13727 and GB30480, and from a DNA marker sequence which we used as an RNAi control. The *fem* and *csd* siRNAs were synthesized (MWG BioTech) and injected at a concentration of 50–100 pg per embryo. Sequences for the nonsense siRNAs used in the control experiments were obtained by scrambling the nucleotide composition of *fem* and *csd* siRNA sequences. Injection of these mock siRNAs did not affect sexual differentiation (data not shown). Individuals derived from inbred crosses (producing diploid males and females) were subject to *csd* genotyping after phenotype analysis at the adult and pupal stage. Head development differs substantially between females (workers) and males (Fig. 1b). Females (workers) have a triangular shaped head, narrow oval-shaped compound eyes, a long proboscis and 12 antennal segments, whereas haploid and diploid males have a round-shaped head, large oval-shaped compound eyes that nearly join in a medio–dorsal position, a short proboscis and 13 antennal segments. Sex-specific splicing of *fem* transcripts were analysed by RT–PCR reactions with RNA prepared from the fourth instar larvae. Amplified fragments were composed of exons 3–6–7–8 (size ~350 bp) and exons 3–4–5–6–7–8 (size ~1.6 kb) corresponding to the female and male transcripts, respectively. Hatched larvae were reared *in vitro*. To quantify the mRNA levels with a BioRad Chromo4, aliquots of first stranded cDNA were amplified, and real-time fluorimetric intensity of SYBR green was monitored. Each sample was run twice in triple replicates. To determine relative transcriptional levels of *csd*, 2 was raised to the power of ΔCts values that were obtained by comparing cycle thresholds (Cts) to those of the reference gene, *elongation factor 1-alpha* ($\Delta Cts = Cts_{control} - Cts_{target}$). Individual *fem* knockdown embryos were identified by testing individual transcription levels against the distribution of mock-siRNA-treated control samples using *t*-test statistics. Multiple comparisons against the distribution were adjusted using the Bonferroni procedure. Statistical analysis was carried out using the SPSS 15.0 software. Sequences of oligonucleotide primers and siRNAs are listed in the Supplementary Information.

Phylogenetic and molecular evolutionary analysis of nucleotide and amino acid substitutions. The proposed phylogenetic tree shows the evolutionary relationship and divergence times of species under study and is based on amber fossils and morphological data. Sequences were obtained from cloned high-fidelity PCR fragments of cDNAs prepared from embryonic mRNA. Extra primer sets to amplify *csd* from *A. cerana* and *A. dorsata* samples were: csd_forCer2/ csd_rev4CIII/; csd_IIIfor/ csd_IIIrev3 (Supplementary Information). The *csd* (*A. mellifera*, $n = 15$; *A. cerana*, $n = 17$; *A. dorsata*, $n = 19$) and *fem* sequences (lacking the Asn/Tyr repeat (hypervariable region); *A. mellifera*, $n = 4$; *A. cerana*, $n = 2$; *A. dorsata*, $n = 2$; *M. compressipes*, $n = 1$; *B. terrestris*, $n = 1$) were aligned by the BioEdit program and edited manually to improve the conformity with the open reading frame. Gaps in the sequence alignments were completely deleted in the evolutionary analyses. Trees of coding sequences were obtained by the minimum evolution method. We applied the Nei–Gojobori distance with Jukes–Cantor correction for synonymous differences and Poisson-corrected distances for amino acid differences which are implemented in the MEGA Version 3.1 program. Gene tree analysis including the wasp *N. vitripennis* was confined to the C- and N-terminal parts of the Fem protein, the region for which we detected sufficient homology (amino acid positions 1–54 and 301–404, XP_001604794). Relative rate tests for substitution rate differences were analysed using the RRTree version 1.1 program on synonymous and amino acid differences using either the *B. terrestris* or the *N. vitripennis fem* sequences as an out-group. Non-synonymous and synonymous substitution analysis of branches of the gene tree were obtained by least square method implemented in the $b_N - b_S$ program. The analysis included up to six *csd* alleles per honeybee species. Fisher's exact test for selection was performed on the numbers of changed and unchanged non-synonymous and synonymous sites. The number of potential non-synonymous and synonymous sites in the branch analysis are 647 and 277, respectively. For Southern blotting onto Hybond N+ membranes standard procedures were used.

LETTERS

Innate immunity induced by composition-dependent RIG-I recognition of hepatitis C virus RNA

Takeshi Saito[1], David M. Owen[1,2], Fuguo Jiang[3], Joseph Marcotrigiano[3] & Michael Gale Jr[1]

Innate immune defences are essential for the control of virus infection and are triggered through host recognition of viral macromolecular motifs known as pathogen-associated molecular patterns (PAMPs)[1]. Hepatitis C virus (HCV) is an RNA virus that replicates in the liver, and infects 200 million people worldwide[2]. Infection is regulated by hepatic immune defences triggered by the cellular RIG-I helicase. RIG-I binds PAMP RNA and signals interferon regulatory factor 3 activation to induce the expression of interferon-α/β and antiviral/interferon-stimulated genes (ISGs) that limit infection[3–10]. Here we identify the polyuridine motif of the HCV genome 3′ non-translated region and its replication intermediate as the PAMP substrate of RIG-I, and show that this and similar homopolyuridine or homopolyriboadenine motifs present in the genomes of RNA viruses are the chief feature of RIG-I recognition and immune triggering in human and murine cells[8]. 5′ terminal triphosphate on the PAMP RNA was necessary but not sufficient for RIG-I binding, which was primarily dependent on homopolymeric ribonucleotide composition, linear structure and length. The HCV PAMP RNA stimulated RIG-I-dependent signalling to induce a hepatic innate immune response *in vivo*, and triggered interferon and ISG expression to suppress HCV infection *in vitro*. These results provide a conceptual advance by defining specific homopolymeric RNA motifs within the genome of HCV and other RNA viruses as the PAMP substrate of RIG-I, and demonstrate immunogenic features of the PAMP–RIG-I interaction that could be used as an immune adjuvant for vaccine and immunotherapy approaches.

To determine the nature of the HCV PAMP RNA, we conducted a functional screen to identify HCV PAMP RNA motifs. We assessed the ability of 1 μg of HCV genome RNA or contiguous subgenomic segments to trigger the interferon (IFN)-β promoter in transfected human Huh7 cells. The full-length HCV genome triggered innate immune signalling to induce the IFN-β promoter (Fig. 1a). Two regions of the HCV RNA, encoding nucleotides 2406–3256 and 8872–9616, significantly induced the IFN-β promoter (Fig. 1b), with signalling activity respectively localized to nucleotides 2406–2696 of the open reading frame and 9389–9619 encoding the 3′ non-translated region (NTR) (Fig. 1c). Deletion of the 3′ NTR but not nucleotides 2408–2663 from the HCV genome significantly attenuated promoter signalling (Fig. 1d). PAMP motifs are typically conserved among strains of a pathogen[1], and sequence comparison of multiple HCV genomes revealed global variability within nucleotides 2406–2696 among virus strains, but nucleotides 9389–9616 encoded motifs of high conservation (Supplementary Fig. 1)[11]. Thus, the viral 3′ NTR might encode HCV PAMP motifs that trigger innate immune signalling in the host cell.

The HCV 3′ NTR is comprised of three regions: a variable region with potential secondary structure; a non-structured poly-U/UC region containing polyuridine with interspersed ribocytidine; and the terminal X region containing three conserved stem-loop structures (Fig. 1e)[12]. We evaluated the ability of RNA encoding the HCV 3′ NTR or each of its regions to trigger intracellular signalling. Because HCV RNA replicates through a negative-sense replication intermediate[2,13], we included analyses of signalling triggered by the replication intermediate counterparts of the 3′ NTR and its composite regions. RNA encoding the genomic or replication intermediate poly-U/UC region was sufficient to trigger signalling to the IFN-β promoter, but neither the variable nor X region genomic and replication intermediate RNA induced promoter signalling (Fig. 1f). The HCV 3′ NTR and poly-U/UC, but not the X region, RNA motifs similarly stimulated signalling when introduced into HeLa cells (Supplementary Fig. 2a). Moreover, in Huh7 cells genomic or replication intermediate 3′ NTR and poly-U/UC RNA each stimulated the formation of active interferon regulatory factor 3 (IRF3) dimers and expression of ISG56, an IRF3 target gene[3], but X region RNA failed to trigger either (Fig. 1g). The poly-U/UC, but not X region, RNA formed a stable complex with purified RIG-I (Fig. 1h). These results define the 100-nucleotide poly-U/UC region of the HCV genome and replication intermediate RNA as the HCV PAMP motif and potential substrate of RIG-I signalling. We also found that the entire HCV 5′ NTR, which contains four major stem-loop structures comprising the viral internal ribosome entry site[14], was only a weak inducer of promoter signalling. However, prior treatment of Huh7 cells with IFN-β to increase RIG-I levels[4] rendered them responsive to signalling triggered by the HCV 5′ NTR or X region RNA (Supplementary Fig. 3). Thus, double-stranded (ds)RNA regions of the HCV RNA are not potent PAMPs but may confer signalling during the IFN response.

To determine the role of RIG-I or other pathogen recognition receptor (PRR) pathways in HCV PAMP signalling, we first examined IFN-β promoter induction in Huh7.5 cells encoding non-functional RIG-I[4]. The cells were refractory to HCV RNA-induced signalling whereas their response was rescued and enhanced on overexpression of wild-type RIG-I (Fig. 2a). MDA5 is a PRR related to RIG-I that binds to dsRNA[15], whereas MyD88 and TRIF are essential adaptor proteins used by Toll-like receptor (TLR) 7/8 and TLR3, respectively, which are PRRs that recognize endosomal single-stranded poly-U RNA or dsRNA[1,16]. We examined PAMP signalling in mouse embryo fibroblasts (MEFs) lacking RIG-I, MDA5, MyD88 or TRIF (Figs 2b–e, Supplementary Fig. 4). When introduced into *RIG-I*$^{-/-}$ MEFs the HCV RNAs did not trigger promoter activation, but 3′ NTR and poly-U/UC RNA, but not X region RNA, stimulated signalling in wild-type, *MDA5*$^{-/-}$, *Myd88*$^{-/-}$ or *Trif*$^{-/-}$ cells. In Huh7 cells, poly-U/UC RNA co-localized and mediated a specific interaction with RIG-I (Fig. 2f). Thus, RIG-I is the essential PRR that signals the innate immune response triggered by HCV poly-U/UC

[1]Department of Immunology, University of Washington School of Medicine, Seattle, Washington 98195-7650, USA. [2]Department of Microbiology, UT Southwestern Medical Center, Dallas, Texas 75235-9048, USA. [3]Department of Chemistry and Chemical Biology, Rutgers University, Piscataway, New Jersey 08854, USA.

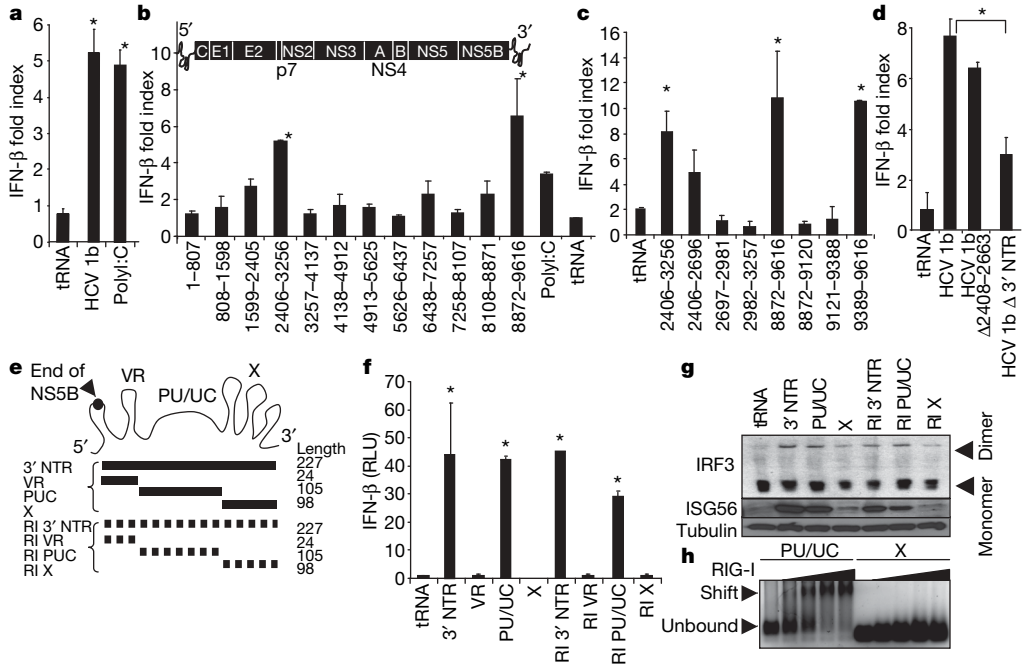


Figure 1 | Identification of HCV PAMP RNA. **a–d**, RNA-induced IFN-β promoter luciferase activity in Huh7 cells, shown as mean fold index induction (compared to non-treated cells; ±s.d.). Huh7 cells were transfected with 1 μg (0.4 pmol) of HCV N (HCV 1b) genome RNA, 1 μg of poly inosine:cytosine (polyI:C) RNA (control) or with 1 μg of the indicated RNA species and harvested for dual luciferase assay 16 h later. HCV 1b refers to HCV genome RNA; tRNA, transfer RNA control. Nucleotide numbers encoded by HCV RNA constructs are shown in **b–d**. Bars are placed in their relative positions of each region within the HCV genome shown in **b**. The 5′ NTR, protein coding regions and 3′ NTR are indicated. **e**, The HCV 3′ NTR motifs and respective RNA constructs. RI and broken lines denote replication intermediate. PUC, PU/UC; VR, variable region. **f**, IFN-β promoter activation, shown here and in the other figures as mean relative luciferase units (RLU; ±s.d.), triggered by 1 μg of the indicated RNA species in transfected Huh7 cells. **g**, The abundance of IRF3, ISG56 and tubulin (control) was measured by immunoblot. The upper panel shows the active IRF3 dimer and inactive monomer forms separated by non-denaturing PAGE. **h**, RNA binding/gel-shift analysis of purified RIG-I with poly-U/UC or X region RNA (6 pmol) reacted with 0, 10, 20, 40, or 60 pmol of RIG-I protein. All RNAs contain 5′ppp. Asterisks indicate significant difference ($P < 0.01$) as determined by Student's *t*-test.

RNA independently of MDA5-, MyD88- or TRIF-dependent PRR pathways.

RIG-I binds to PAMP RNA containing 5′ terminal triphosphate (5′ppp) through which the triphosphate end is proposed to anchor the RNA within charged residues of the RIG-I repressor domain, causing a conformation change to displace the repressor domain and release signalling autorepression[17–19]. Gel-shift assays revealed that 5′ppp was required for poly-U/UC RNA binding by RIG-I but did not mediate stable RIG-I interaction with X region RNA (Fig. 3a). 5′ppp was required for IFN-β promoter signalling by poly-U/UC

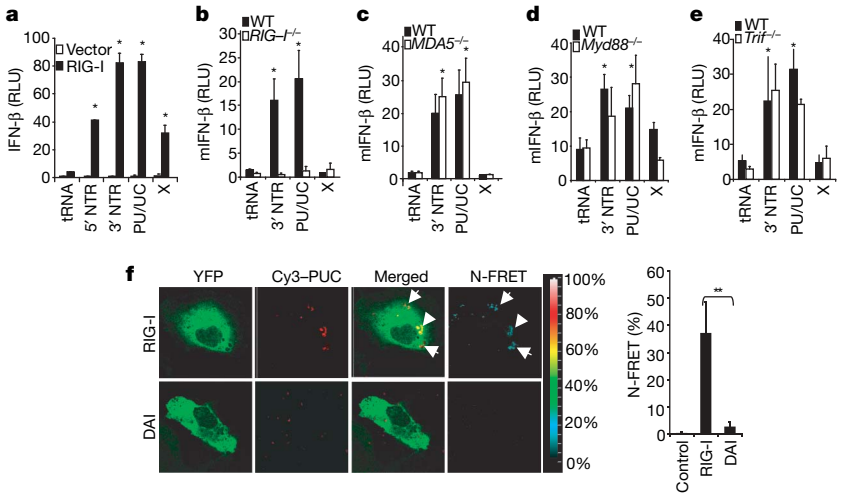


Figure 2 | RIG-I-specific HCV RNA PAMP recognition and signalling. **a–e**, Induction of the IFN-β promoter in cells co-transfected with 1 μg of tRNA control or the indicated HCV RNA species. **a**, Huh7.5 cells, lacking functional RIG-I, were co-transfected with a plasmid encoding vector alone or RIG-I. Promoter signalling in wild-type (WT) and *RIG-I*$^{-/-}$ (**b**), *MDA5*$^{-/-}$ (**c**), *Myd88*$^{-/-}$ (**d**) or *Trif*$^{-/-}$ (**e**) MEFs. Asterisks indicate a significant difference ($P < 0.01$) from tRNA control. Similar results were obtained when cells were transfected with 30 pmol of each RNA (data not shown). **f**, FRET analysis of Cy3-labelled poly-U/UC RNA (Cy3–PUC) interaction with YFP–RIG-I or YFP–DAI protein in co-transfected Huh7 cells. Panels show representative images of YFP, Cy3, merged fluorescence and N-FRET (corrected FRET). The colour scale denotes N-FRET levels. The bar graph at the right shows the calculated values for RNA interaction with RIG-I or DAI (% N-FRET; ±s.d.). Control values are from the image area that has no co-localization signal. All RNAs contain 5′ppp.

RNA, and supported low-level promoter induction triggered by X region RNA in IFN-treated cells (Fig. 3b). Because X region RNA failed to form a stable complex with RIG-I and only weakly triggered signalling, stable RIG-I–RNA interaction is probably required to release RIG-I autorepression. We therefore conducted limited trypsin digestion analysis of purified RIG-I alone or bound to poly-U/UC or X region RNA containing 5′ppp. This approach provides an assessment of RIG-I repressor domain displacement in response to PAMP RNA binding wherein the displaced repressor domain of signalling-active RIG-I presents as a protected 17-kDa fragment[17–19]. As shown in Fig. 3c, RIG-I binding of poly-U/UC but not X region RNA rendered the protected 17-kDa repressor domain fragment. These results demonstrate that 5′ppp is necessary but not sufficient for RNA binding by RIG-I wherein the HCV poly-U/UC RNA directs stable interaction with RIG-I in a 5′ppp-dependent manner that confers signalling activation. We used 5′ppp RNA in all further experiments.

The HCV polyU/UC region is a flexible motif among HCV strains and is essential for viral replication[13]. In the HCV genotype 1b strain used in these experiments, the poly-U/UC region is comprised of 100 nucleotides containing 78% uridine and 22% ribocytosine (Supplementary Fig. 5). A reduction of length through progressive 3′ truncation to 50 nucleotides or fewer attenuated signalling in Huh7 cells (Fig. 3d), consistent with reduced RIG-I binding of short RNAs[20]. Replacement of ribocytosine with riboguanine (poly-U/UG) had no impact on PAMP signalling, but progressive replacement of uridine for riboguanine to below 80% uridine or 100% polyriboguanine (poly-G) reduced or abrogated PAMP signalling (Fig. 3e). Because the replication intermediate poly-U/UC RNA contains high polymeric riboadenine content, we also examined the impact of poly-A composition and length on RIG-I signalling. One-hundred-nucleotide poly-A RNA and poly-U/UC RNA equally induced signalling to the IFN-β promoter, whereas reduced riboadenine content of the former through progressive nucleotide replacement to poly-G attenuated or ablated signalling (Fig. 3f). In side-by-side analyses we found that poly-U/UC, poly-U/UG and poly-A RNA, but neither poly-C nor poly-G RNA, could trigger signalling to the IFN-β promoter in Huh7 (Fig. 3g) and HeLa cells (Supplementary Fig. 2b). Truncation of the poly-A RNA to 50 nucleotides or fewer attenuated signalling to the IFN-β promoter to the same extent as truncation of poly-U/UC (Supplementary Fig. 2c). Whereas poly-C and poly-G RNA bound negligibly to RIG-I, poly-U/UC, poly-U/UG and poly-A RNA formed a stable complex with RIG-I (Supplementary Fig. 5e) that released the RIG-I repressor domain to the active conformation (Fig. 3h). These results define polymeric uridine or riboadenine motifs of 50 nucleotides or greater—including the PU/UC and replication intermediate PU/UC motif of HCV—as the PAMP signature within 5′ppp RNA that is efficiently recognized by RIG-I to trigger the immune response.

To determine whether RIG-I recognition of the HCV poly-U/UC PAMP motif triggers hepatic innate immune defences *in vivo*, we conducted RNA signalling analysis in wild-type and *RIG-I*$^{-/-}$ mice. Intravenous administration of full-length HCV 1b genome stimulated hepatic *Ifnb* messenger RNA expression within 8 h in wild-type mice but not in *RIG-I*$^{-/-}$ mice, and signalling was significantly attenuated on deletion of the HCV 3′ NTR (Fig. 4a). Moreover, the poly-U/UC RNA motif, but not the X region RNA motif, was sufficient to trigger the hepatic IFN-β expression in wild-type but not in *RIG-I*$^{-/-}$ mice. In time course studies we found that the poly-U/UC RNA motif

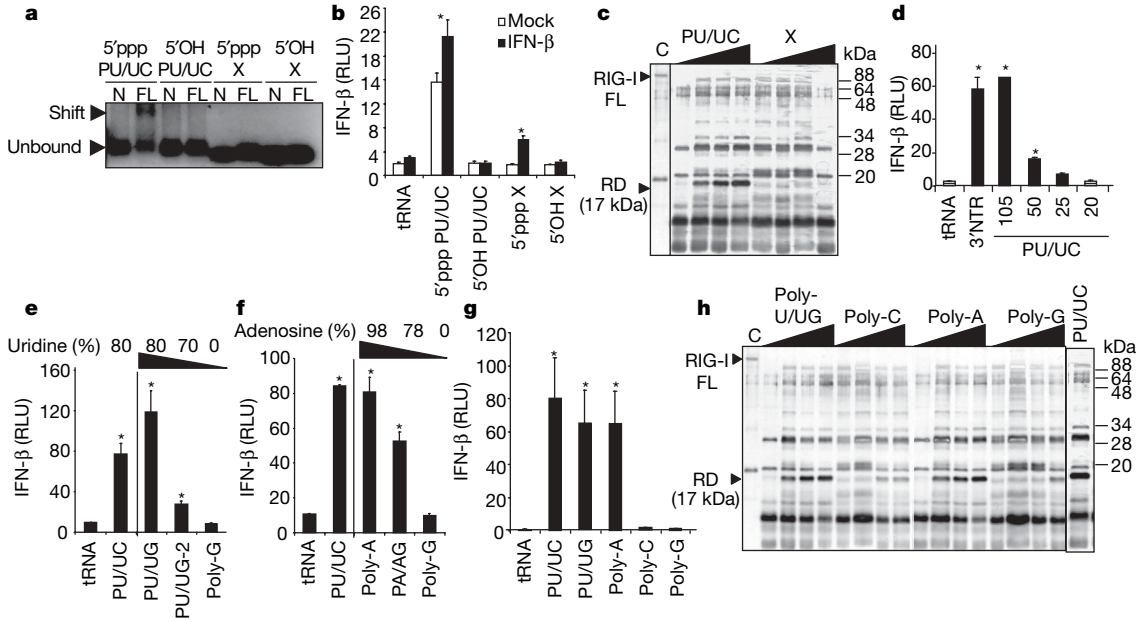


Figure 3 | Polyuridine and polyriboadenine ribonucleotides are RIG-I ligands. **a**, Gel-shift analysis of complex formation between 25 pmol of purified N-RIG (RIG-I amino acids 1–228, control) or full-length RIG-I (FL) and 10 pmol of poly-U/UC (PU/UC) or X region RNA containing 5′ppp or 5′OH as indicated. Arrows denote position of unbound RNA and RNA–RIG-I complexes. **b**, Effect of 5′ppp on IFN-β promoter activity. Huh7 cells were either mock-treated or treated with IFN-β 8 h before transfection with 1 μg (30 pmol) of RNA. **c**, Effect of poly-U/UC or X region RNA on RIG-I activation. The silver-stained gel image shows trypsin digestion products of RIG-I that was pre-incubated with increasing amounts poly-U/UC or X region RNA. Arrows indicate positions of full-length (FL) RIG-I and the 17-kDa trypsin-resistant repressor domain (RD) from RIG-I–RNA complexes. **d**, Effect of nucleotide length of 1 μg poly-U/UC 3′ truncation products on IFN-β promoter signalling in Huh7 cells. IFN-β promoter activation is shown as mean relative luciferase units (RLU; ±s.d.). Similar results were obtained when cells were transfected with 30 pmol of each RNA (data not shown). **e–g**, Effect of nucleotide composition on IFN-β promoter signalling in Huh7 cells transfected with 1 μg (30 pmol) of RNA. **h**, Effect of nucleotide composition on RIG-I activation. The silver-stained gel image shows trypsin digestion products of RIG-I that was pre-incubated with increasing amounts poly-U/UG, poly-C, poly-A, or poly-G RNA. Arrows indicate positions of full-length RIG-I and the 17-kDa trypsin-resistant repressor domain. We confirmed the 17-kDa fragment as the RIG-I repressor domain by immunoblot analysis of the digestion products using an antiserum specific to the RIG-I carboxy terminus (not shown), as previously described[17]. Asterisks indicate significant difference ($P < 0.01$) as determined by Student's *t*-test.

induced a peak of hepatic *Ifnb* mRNA expression and IFN-β serum levels at 8 h after injection in wild-type mice (Fig. 4b, c). This response was associated with induced hepatic expression of *RIG-I* and *ISG56* mRNA and tissue-wide expression of hepatic ISG54 (Fig. 4d–f), similar to the hepatic response in HCV-infected patients[9,10]. *RIG-I*$^{-/-}$ mice expressed only a low level of *Ifnb* and *ISG56*. The tissue-wide nature of hepatic ISG54 expression in wild-type mice suggests that paracrine signalling of IFN-β could have an important role in hepatic defences against HCV. To test this idea we measured HCV production in infected Huh7 cells that were treated with IFN-β or supernatants collected from cultures transfected with HCV poly-U/UC RNA, X-region RNA, or tRNA (control). Poly-U/UC RNA triggered IFN-β expression in the transfected cells (data not shown), and only treatment with IFN-β or supernatant from the poly-U/UC-transfected cells induced a response that suppressed HCV infection (Fig. 4g, h). Thus, the poly-U/UC RNA is an HCV genome PAMP that is necessary and sufficient to trigger RIG-I signalling of the hepatic innate immune response. The actions of RIG-I signalling can induce an antiviral response directly (Fig. 1g), as well as through indirect, paracrine actions of IFN produced from HCV PAMP signalling (Fig. 4g, h).

Our results provide new insights into the features of PAMP specificity of RIG-I wherein RNA virus genome sequences consisting of poly-U and respective replication intermediate poly-A motifs of length >50 nucleotides are the determinants that confer efficient RIG-I binding and signalling. 5′ppp was necessary but not sufficient for stable binding of HCV PAMP RNA by RIG-I. In terms of the HCV genome, well defined internal RNA interactions of the 5′ and 3′ ends[21,22] could provide 5′ppp and PAMP motif proximity for stable RIG-I binding. The poly-U/UC motif is an essential determinant of HCV replication fitness[21]. Thus, as the virus must maintain this motif for its viability, the host takes advantage of this requirement and targets the poly-U/UC region as a discriminator of PAMP RNA through RIG-I interaction. Poly-U and/or poly-A motifs are present in localized regions in the genome of RNA viruses known to trigger RIG-I signalling (Supplementary Table 1)[5,8,19]. We found that 5′ppp genomic poly-U/A-rich RNA motifs within the rabies virus leader sequence, Ebola virus 3′ region, or the measles virus leader sequence each triggered signalling to the IFN-β promoter in Huh7 cells, but GC-rich RNA motifs from each viral genome did not trigger a significant response (Supplementary Fig. 7a, b). We also found that pppT3-63, Tri-GFP and EGFP 2 T7 RNAs, previously described as RIG-I substrates and comprised of 50% or fewer A/U nucleotides[5,18,23], could induce only weak signalling to the IFN-β promoter compared to the poly-U/UC RNA . Thus, A/U composition and poly-U motifs are the major determinants of viral PAMP RNA recognition by RIG-I. Cellular RNAs also contain poly-U and poly-A motifs but mRNAs are typically capped and are bound by poly-A binding proteins[24], whereas ribosomal RNAs are 'masked' as ribonucleoprotein complexes[25]. These features and the context of 5′ppp with viral poly-U and poly-A motifs serve to identify self from non-self RNA by governing RIG-I recognition, wherein, non-self recognition of the HCV PAMP RNA triggers a hepatic innate immune response. These observations provide a possible explanation of why 25% or more of all HCV-exposed people clear acute infection[2] and why HCV needs to evade innate immunity through viral NS3/4A protease targeting of the RIG-I pathway[26,27]. RIG-I substrates such as the poly-U/UC RNA or structurally similar compounds could provide therapeutic application as immune adjuvants similar to TLR agonists[28], and offer innate immune stimulatory properties that

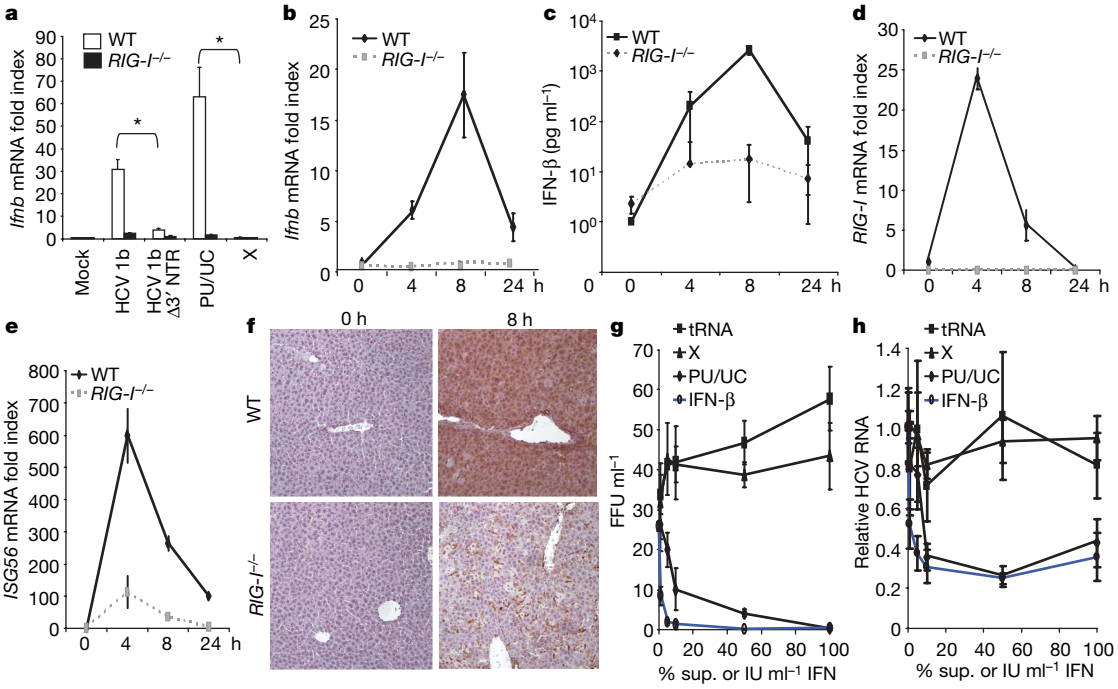


Figure 4 | HCV PAMP RNA triggers the hepatic innate immune response and anti-HCV defences. **a**–**f**, Wild-type or *RIG-I*$^{-/-}$ mice ($n = 3$) were hydrodynamically transfected intravenously with HCV RNA. **a**, Mice received 100 μg of HCV 1b genome, HCV 1b genome lacking the 3′ NTR (HCV 1b Δ3′ NTR), PU/UC RNA or X region RNA. Hepatic *Ifnb* mRNA expression was measured 8 h later. **b**–**f**, Wild-type or *RIG-I*$^{-/-}$ mice ($n = 3$) received 200 μg of poly-U/UC RNA or buffer control, and were killed 4, 8 or 24 h later for comparative measurement of mRNA and protein expression. **b**, Liver-specific expression of *Ifnb* mRNA. **c**, Serum IFN-β protein levels. **d**, Liver-specific expression of *RIG-I* mRNA. **e**, Liver-specific expression of *ISG56* mRNA. **f**, Immunohistochemical stain of ISG54 protein expression in liver tissue sections. **g**, **h**, Paracrine antiviral effect of the innate immune response triggered by HCV PAMP RNA. **g**, Inhibition of HCV infection in pre-treated cells. Triplicate cultures of Huh7.5 cells were treated with DMEM containing increasing concentrations of IFN-β or conditioned media collected from Huh7 cells transfected with the indicated RNA species for 12 h before HCV infection. The graph shows the number of infected cells (±s.d.) as determined by focus-forming unit (FFU) assay at 48 h after infection. **h**, Huh7.5 cells were infected with HCV for 48 h and then were treated with increasing international units per ml concentrations of IFN-β (IU ml^{-1} IFN) or the indicated amount of conditioned media (% sup.) for an additional 48 h. Intracellular HCV RNA levels relative to *GAPDH* were determined and are plotted as mean HCV RNA index (±s.d.) relative to infected, untreated cells.

may improve IFN-based therapy for HCV through paracrine immune actions that limit infection[2].

METHODS SUMMARY

RNA. RNA constructs and quality control analysis are shown in Supplementary Fig. 5. 5′ppp RNA was synthesized using the T7 Megascript kit (Ambion). Full-length and subgenomic HCV RNA were produced from plasmid DNA or T7 promoter-linked PCR products generated from cloned HCV N (A gift from S. Lemon) or Con1 genome (HCV genotype-1b)[29]. 5′OH RNAs were purchased from Fidelity Systems. RNA transfection was performed using 1 μg of RNA per 1×10^5 cells with the Transmessenger reagent (Qiagen). One microgram of RNA mass approximates to the following picomoles: full-length HCV 1b genome, HCV 1b Δ2408–2663 and HCV 1b Δ3′ NTR, 0.4 pmol; HCV 1b subgenomic RNA constructs, 5 or 15 pmol; HCV genotype 1b (Con1) 5′ NTR, 3′ NTR (20 pmol), and all other RNA constructs, 30–150 pmol. RNA concentrations in the transfection mix were 5–10 μg ml^{-1}. Additionally, in experiments to assess signalling induced by subgenomic HCV RNA, we also measured promoter expression triggered by equimolar amounts of RNA. RNA delivery was assessed as described in Supplementary Fig. 6. mRNA expression was determined by quantitative polymerase chain reaction with reverse transcription assay. DNA oligonucleotides used in this study are described in Supplementary Table 2.

RNA signalling analysis. IFN-β promoter luciferase analyses of transfected cells were conducted as described[17]. Protein expression and the abundance of IRF3 dimer and monomeric forms in cells were determined by immunoblot analysis[17].

RIG-I purification and RNA binding analysis. Full-length RIG-I or RIG-I amino acids 1–228 (N-RIG) were expressed in *Escherichia coli* and purified. RIG-I–RNA complexes were assessed by gel-shift assay and SYBR green staining (Lonza). RIG-I activation/conformation shift was analysed using the limited trypsin digestion method[17]. FRET analysis of Cy3-poly-U/UC RNA interaction with YFP–RIG-I or YFP–DAI proteins was conducted using N-FRET on a Zeiss confocal microscope[30]. Serum IFN-β levels were measured by ELISA (PBL).

Mice. Mice[15] were from S. Akira[15] and were transfected using lipid-based *in vivo* RNA transfection reagent (Altogen).

Statistical analysis. Data were compared using the Student's *t*-test.

Received 13 February; accepted 23 May 2008.
Published online 11 June 2008.

1. Saito, T. & Gale, M. Principles of intracellular viral recognition. *Curr. Opin. Immunol.* **19,** 17–23 (2007).
2. Lauer, G. M. & Walker, B. D. Medical progress: Hepatitis C virus infection. *N. Engl. J. Med.* **345,** 41–52 (2001).
3. Gale, M. & Foy, E. M. Evasion of intracellular host defence by hepatitis C virus. *Nature* **436,** 939–945 (2005).
4. Sumpter, R. *et al.* Regulating intracellular antiviral defense and permissiveness to hepatitis C virus RNA replication through a cellular RNA helicase, RIG-I. *J. Virol.* **79,** 2689–2699 (2005).
5. Hornung, V. *et al.* 5′-triphosphate RNA is the ligand for RIG-I. *Science* **314,** 994–997 (2006).
6. Pichlmair, A. *et al.* RIG-I-mediated antiviral responses to single-stranded RNA bearing 5′-phosphates. *Science* **314,** 997–1001 (2006).
7. Yoneyama, M. *et al.* The RNA helicase RIG-I has an essential function in double-stranded RNA-induced innate antiviral responses. *Nature Immunol.* **5,** 730–737 (2004).
8. Loo, Y. M. *et al.* Distinct RIG-I and MDA5 signaling regulation by RNA viruses in innate immunity. *J. Virol.* **27,** 697 (2007).
9. Lau, D. T. *et al.* Interferon regulatory factor-3 activation, hepatic interferon-stimulated gene expression, and immune cell infiltration in hepatitis C virus patients. *Hepatology* **47,** 799–809 (2008).
10. Smith, M. W. *et al.* Gene expression patterns that correlate with hepatitis C and early progression to fibrosis in liver transplant recipients. *Gastroenterology* **130,** 179–187 (2006).
11. Simmonds, P. Genetic diversity and evolution of hepatitis C virus - 15 years on. *J. Gen. Virol.* **85,** 3173–3188 (2004).
12. Kolykhalov, A. A., Mihalik, K., Feinstone, S. M. & Rice, C. M. Hepatitis C virus-encoded enzymatic activities and conserved RNA elements in the 3′ nontranslated region are essential for virus replication *in vivo*. *J. Virol.* **74,** 2046–2051 (2000).
13. Yi, M. K. & Lemon, S. M. 3′ Nontranslated RNA signals required for replication of hepatitis C virus RNA. *J. Virol.* **77,** 3557–3568 (2003).
14. Honda, M. *et al.* Structural requirements for initiation of translation by internal ribosome entry within genome-length hepatitis C virus RNA. *Virology* **222,** 31–42 (1996).
15. Kato, H. *et al.* Differential roles of MDA5 and RIG-I helicases in the recognition of RNA viruses. *Nature* **441,** 101–105 (2006).
16. Heil, F. *et al.* Species-specific recognition of single-stranded RNA via toll-like receptor 7 and 8. *Science* **303,** 1526–1529 (2004).
17. Saito, T. *et al.* Regulation of innate antiviral defenses through a shared repressor domain in RIG-I and LGP2. *Proc. Natl Acad. Sci. USA* **104,** 582–587 (2007).
18. Takahasi, K. *et al.* Nonself RNA-sensing mechanism of RIG-I helicase and activation of antiviral immune responses. *Mol. Cell* **29,** 428–440 (2008).
19. Cui, S. *et al.* The C-terminal regulatory domain is the RNA 5′-triphosphate sensor of RIG-I. *Mol. Cell* **29,** 169–179 (2008).
20. Marques, J. T. *et al.* A structural basis for discriminating between self and nonself double-stranded RNAs in mammalian cells. *Nature Biotechnol.* **24,** 559–565 (2006).
21. You, S. & Rice, C. M. 3′ RNA elements in hepatitis C virus replication: kissing partners and long poly(U). *J. Virol.* **82,** 184–195 (2008).
22. Ito, T. & Lai, M. M. C. An internal polypyrimidine-tract-binding protein-binding site in the hepatitis C virus RNA attenuates translation, which is relieved by the 3′-untranslated sequence. *Virology* **254,** 288–296 (1999).
23. Kim, D. H. *et al.* Interferon induction by siRNAs and ssRNAs synthesized by phage polymerase. *Nature Biotechnol.* **22,** 321–325 (2004).
24. Afonina, E., Stauber, R. & Pavlakis, G. N. The human Poly(A)-binding protein 1 shuttles between the nucleus and the cytoplasm. *J. Biol. Chem.* **273,** 13015–13021 (1998).
25. Yusupov, M. M. *et al.* Crystal structure of the ribosome at 5.5 Å resolution. *Science* **292,** 883–896 (2001).
26. Meylan, E. *et al.* Cardif is an adaptor protein in the RIG-I antiviral pathway and is targeted by hepatitis C virus. *Nature* **437,** 1167–1172 (2005).
27. Loo, Y. M. *et al.* Viral and therapeutic control of IFN-β promoter stimulator 1 during hepatitis C virus infection. *Proc. Natl Acad. Sci. USA* **103,** 6001–6006 (2006).
28. Tse, K. & Horner, A. A. Update on toll-like receptor-directed therapies for human disease. *Ann. Rheum. Dis.* **66** (suppl. 3), 77–80 (2007).
29. Beard, M. R. *et al.* An infectious molecular clone of a Japanese genotype 1b hepatitis C virus. *Hepatology* **30,** 316–324 (1999).
30. Takaoka, A. *et al.* DAI (DLM-1/ZBP1) is a cytosolic DNA sensor and an activator of innate immune response. *Nature* **448,** 501–505 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank T. Fujita for discussions, and S. Horner for manuscript review. We thank T. Fujita, S. Lemon, G. Sen, C. Rice, T. Taniguchi, A. Miyawaki, S. Akira for reagents, R. Hirai for technical consultation, and J. Briley, S. Thomas and G. Martin for technical assistance. This work was supported by funds from the State of Washington, National Institutes of Health grants R01AI060389, R01DA021353, U19AI40035 (Project 4) and the Burroughs-Wellcome Fund, and by a gift from Mr. and Mrs. R. Batcheldor.

Author Contributions T.S. conducted RNA binding studies and RIG-I signalling analyses. T.S. and D.M.O. conducted *in vivo* studies. F.J. and J.M. developed the RIG-I protein-expression system, and produced, purified and tested recombinant RIG-I proteins. M.G. directed the research. All authors participated in study design and manuscript preparation. T.S. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to M.G. (mgale@u.washington.edu).

LETTERS

Imbalance between pSmad3 and Notch induces CDK inhibitors in old muscle stem cells

Morgan E. Carlson[1], Michael Hsu[1] & Irina M. Conboy[1]

Adult skeletal muscle robustly regenerates throughout an organism's life, but as the muscle ages, its ability to repair diminishes and eventually fails[1,2]. Previous work suggests that the regenerative potential of muscle stem cells (satellite cells) is not triggered in the old muscle because of a decline in Notch activation, and that it can be rejuvenated by forced local activation of Notch[3]. Here we report that, in addition to the loss of Notch activation, old muscle produces excessive transforming growth factor (TGF)-β (but not myostatin), which induces unusually high levels of TGF-β pSmad3 in resident satellite cells and interferes with their regenerative capacity. Importantly, endogenous Notch and pSmad3 antagonize each other in the control of satellite-cell proliferation, such that activation of Notch blocks the TGF-β-dependent upregulation of the cyclin-dependent kinase (CDK) inhibitors p15, p16, p21 and p27, whereas inhibition of Notch induces them. Furthermore, in muscle stem cells, Notch activity determines the binding of pSmad3 to the promoters of these negative regulators of cell-cycle progression. Attenuation of TGF-β/pSmad3 in old, injured muscle restores regeneration to satellite cells *in vivo*. Thus a balance between endogenous pSmad3 and active Notch controls the regenerative competence of muscle stem cells, and deregulation of this balance in the old muscle microniche interferes with regeneration.

Satellite cells are muscle stem cells capable of lifelong maintenance and repair of myofibres, or differentiated muscle cells[4–6]. The decline in muscle tissue regeneration with age is largely due to a decreased activation of Notch pathway, which is required for satellite cells to break quiescence and prevents premature differentiation into myotubes by antagonizing Wnt pathway[3–9]. Forced activation of Notch-1, by antibody specific to its external domain (Notch-1 cleavage and activation), rejuvenates muscle repair; whereas inhibiting Notch in young muscle interferes with regeneration[3], suggesting that Notch pathway is an essential and age-specific molecular determinant of adult myogenesis.

It is widely believed that the microenvironment controls resident stem-cell behaviour[10], and our recent work established that aged muscle fibres inhibit the regenerative responses of muscle stem cells[11]. Here we examine the biochemical changes occurring in the aged microniche of muscle stem cells, for example differentiated myofibres, focusing on the age-specific interplay between active Notch and TGF-β/pSmad3 and on the ability of this signal integration to control levels of CDK inhibitors in satellite cells.

Binding of activated TGF-β proteins to their receptors induces the phosphorylation and activation of the Smad transcription factors, which form heteromers (Smad4 as a common component) that translocate into the nucleus[12–15]. Different ligands, for example TGF-β1, -β2, -β3 and myostatin, are capable of activating the same Smad2, 3 proteins[12,16,17]. Increased TGF-β signalling has been implicated in the inhibition of cell-cycle progression (both generally and in myogenic lineage), by activating CDK inhibitors and inactivating cMyc[14,18–22]. Importantly, recent genetic-targeting experiments suggest that during ageing the necessity to impose cell-cycle checkpoints becomes antagonistic to the regenerative responses of adult stem cells[23–26].

This report establishes the following causalities: (1) old myofibres inhibit their own repair by shifting balance from active Notch to over-pronounced pSmad3 in resident muscle stem cells, which upregulates p15, p16, p21 and p27 and thwarts satellite-cell regenerative capacity; (2) active Notch can override this block of satellite-cell responses by removal of pSmad3 from CDK inhibitor promoters.

Growth factors, including TGF-β family members, are typically localized and activated in a tissue's extracellular matrix (ECM), which in skeletal muscle is the main component of the basement membrane surrounding myofibres and their associated satellite cells[27].

As shown in Fig. 1, there is dramatic and constant upregulation of functional TGF-β, but not the muscle-specific family member myostatin[28], in the aged, injured and resting muscle compared with young. As expected, when present at high levels, TGF-β co-localizes with the laminin^{+} basement membrane (the immediate microniche of muscle stem cells) (Fig. 1a)[29]. Isotype-matched antibody controls for these and other experiments were negative (Supplementary Fig. 1a). These data were confirmed in western blot analysis, which also established that with age both differentiated myofibres and satellite cells, located in resting and injured muscle, upregulate levels of TGF-β (inactive precursor plus bioactive protein) and pSmad3 (Fig. 1b–e and Supplementary Fig. 2). In contrast, levels of myostatin and follistatin are not changed with age (Fig. 1a–c). In agreement with previously published work[3], the levels of active Notch were reciprocal to those of pSmad3 and TGF-β, which is high in young and low in old satellite cells (Fig. 1b and Supplementary Fig. 1b). The purity of satellite cells in these preparations is greater than 95% (Supplementary Fig. 3)[3]. High levels of nuclear pSmad3 were also detected in satellite cells residing in the aged muscle *in vivo* (Supplementary Fig. 4), and excessive TGF-β was detected in old muscle at days 1 and 3 after injury (not shown).

Our recent report demonstrated that proliferation and myogenesis of even young satellite cells are inhibited by the aged myofibres[11]. Interestingly, factors secreted by aged myofibres rapidly upregulated TGF-β production by young satellite cells (Fig. 1f) in transwell co-cultures (impermeable to cell migration). This provides a molecular explanation for the pre-mature 'ageing' of young progenitor cells exposed to aged tissue.

These data establish that although both Notch and pSmad3 can be robustly activated in skeletal muscle, because their ligands are expressed by myofibres and satellite cells, with age the balance is shifted from active Notch to active TGF-β/pSmad3 (Fig. 1)[3].

During the first days after injury, satellite cells need to break quiescence and proliferate. However, there is an age-specific elevation of pSmad3 and diminished Notch activation, either one of which is

[1]Department of Bioengineering, University of California, Berkeley, California 94720, USA.

sufficient to inhibit cell-cycle progression[17,22]. We hypothesized that excessive levels of TGF-β/pSmad might upregulate the levels of CDK inhibitors in muscle stem cells, whereas activation of Notch might antagonize this process. After muscle injury, satellite cells were derived from young muscle[3] and cultured with TGF-β1, with or without simultaneous forced activation of Notch. Compared with untreated cells, exogenously added TGF-β1 caused a prompt upregulation of p15, p16, p21 and p27 in satellite cells (Fig. 2a, quantified in Fig. 2b). When endogenous Notch was experimentally activated, simultaneously with TGF-β1 treatment, the inducing effects on p15, p16, p21 and p27 levels were significantly attenuated (Fig. 2a, b and Supplementary Fig. 5c) at a range of TGF-β1 concentrations (Supplementary Fig. 5). Manipulation of TGF-β/pSmad and active Notch balance not only controls CDK inhibitor levels, but also regulates proliferation of satellite cells in their endogenous microniches (Supplementary Fig. 6).

To analyse this antagonistic interaction with higher precision, we examined whether active Notch and pSmad3 physically interact on promoters of the p15, p16, p21 and p27 genes. A pSmad3-specific chromatin immunoprecipitation assay (ChIP) was performed on satellite cells treated with TGF-β only, activation of Notch only, TGF-β and activation of Notch together or untreated.

As shown in Fig. 2c, both active Notch and RNA polymerase II are detected in a complex with pSmad3, suggesting that active Notch and pSmad3 physically interact on gene regulatory regions. Consistent with the idea of functional balance, forced activation of Notch yielded more endogenous Notch and treatment with TGF-β yielded more pSmad3 in these complexes (Fig. 2c).

DNA co-precipitated with pSmad3 was analysed by quantitative polymerase chain reaction (Q-PCR), using primers specific for 5′ promoter regulatory regions of p15, p16, p21 and p27 (Fig. 2d, e). These data provided a further understanding of the molecular mechanism of active Notch and TGF-β/pSmad3 antagonism. Namely, forced activation of Notch dramatically reduced pSmad3 presence on the promoter regions of p15, p16, p21 and p27, even in the presence of TGF-β treatment (Fig. 2d, e).

Interestingly, in the absence of Notch activation (by γ-secretase inhibitor (GSI)), young/low levels of TGF-β are sufficient to induce p15, p16, p21 and p27 proteins (Fig. 3a, b); and pSmad3 presence on the promoters of these genes increases in young muscle stem cells (Fig. 3d, e). Expectedly, less active Notch was detected in Notch–pSmad3 complexes after treatment with GSI, whereas pSmad3 levels were not affected (Fig. 3c). These results were confirmed with several sets of PCR primers spanning about 1 kilobase of the studied regulatory regions; and corroborated by negative controls, including the lack of amplification of a Smad3 non-enriched genome region (Supplementary Fig. 7).

These data show that in muscle stem cells, active Notch attenuates age-specific induction of multiple CDK inhibitors by reducing the presence of pSmad3 on their promoter regions; and that either diminished Notch activation or increased TGF-β/pSmad3 levels are sufficient for upregulating these negative regulators of cell-cycle progression (Figs 2 and 3, and Supplementary Figs 6 and 7). Interestingly, p16 is sensitive to a slight increase in pSmad3, as inhibition of Notch robustly enhanced pSmad3 binding, whereas Notch activation did not significantly reduce TGF-β imposed pSmad3 binding to the p16 promoter (Figs 2, 3 and Supplementary Fig. 7).

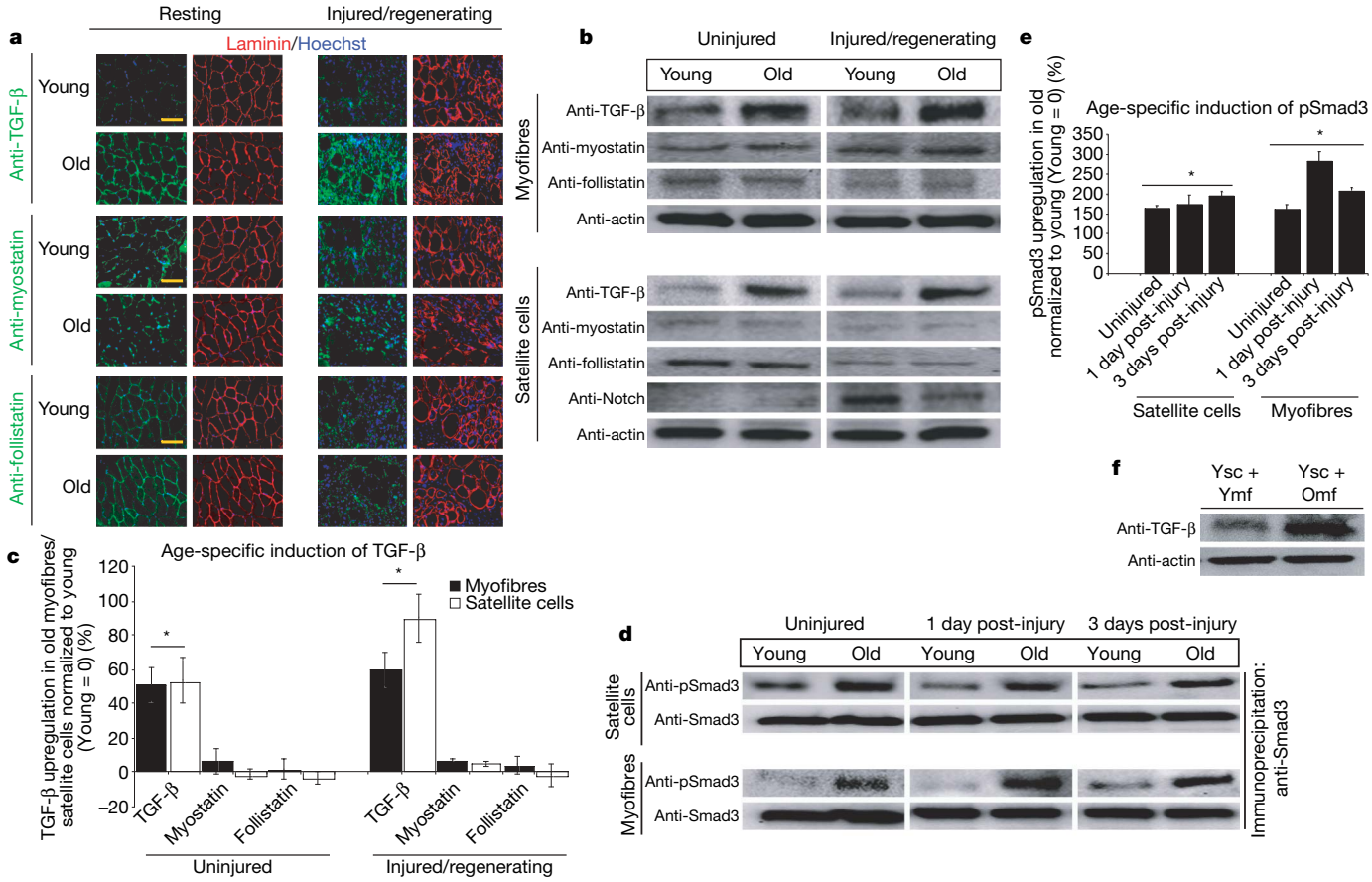


Figure 1 | TGF-β/pSmad3, but not myostatin, increases in old skeletal muscle. **a**, Immunodetection of TGF-β, myostatin or follistatin (green) and laminin (red) in 10-μm skeletal muscle cryosections. Hoechst labels nuclei (blue). Scale bar, 50 μm. **b**, Western blot on myofibres and satellite cells; quantified in **c**. **d**, Immunoprecipitation with anti-Smad3 antibody, followed by western blot with anti-phosphorylated Smad3 antibody; quantified in **e**. **f**, After overnight transwell co-culture with young or old myofibres, young satellite cells were analysed by western blotting; data are means ± s.d., $n = 3$. *$P \leq 0.05$ compared with young.

To confirm these findings and address their physiological significance *in vivo*, we examined whether aged muscle stem cells would effectively repair old tissue when pSmad3 is locally attenuated by lentivirally delivered short hairpin RNA (shRNA) to *Smad3*.

Young and old muscles were infected with the following lentiviral particles *in vivo*: three different *Smad3*-targeted shRNAs, non-target shRNA to control for non-specific RNA interference or transduction control; and followed by muscle injury (Fig. 4 and Supplementary Figs 8–10). As expected, at 5 days after injury, old control muscle contains fewer regenerated myofibres than young, based on haematoxylin and eosin histology and immunodetection of eMyHC *de novo* myofibres with centrally located bromodeoxyuridine ($BrdU^+$) nuclei that are the fusion product of myoblasts generated by satellite cells (Fig. 4)[3]. In contrast to control viral transduction and non-target shRNA control, the acute *in vivo* expression of each one of the *Smad3*-targeted shRNAs enhanced and rejuvenated regeneration of old muscle (Fig. 4a, b and Supplementary Fig. 10a). Furthermore, in old muscle satellite cells there was pronounced increase in p15 (at all times) and p21 (upon muscle injury); and importantly, p15 and p21 were attenuated *in vivo* to their 'youthful' levels by each tested *Smad3*-targeted shRNA, but not by non-target shRNA (Fig. 4c, d). The levels of p16 and of p21, p27 were not increased in quiescent satellite cells endogenous to old resting muscle, consistent with rapid activation of their myogenic responses in young environments (Fig. 4c, d)[3,11]. A slight increase in p16 and p27 levels was detected 24 h after injury, suggesting a potential physiological importance for early muscle stem-cell responses (Supplementary Fig. 9). Thus, elevated TGF-β/pSmad3 assures the age-dependent upregulation of at least one CDK inhibitor at all times, while specific time points are expectedly different for individual CDK inhibitors known to be under control of many molecular cues. These data strongly suggest that Smad3-specific (rather than non-target effects[30]) rescued the satellite-cell regenerative responses in old niches, as three different *Smad3*-targeted shRNAs similarly enhanced myogenesis of old muscle and diminished the levels of p15 and p21, but not of p16 and p27 in the aged satellite cells (Fig. 4a–d and Supplementary Fig. 10). *Smad3* targeting of these shRNAs is confirmed by the expected downregulation of messenger RNA levels of *Smad3* and nuclear levels of pSmad3 protein, as well as of Smad3, Smad6, Smad7, TGF-β and myostatin in satellite cells *in vivo*[14,15] (Supplementary Fig. 10b–e). Notably, pan-neutralizing antibody against TGF-β rejuvenated repair of old muscle, and recombinant TGF-β resulted in scarring of young muscle (Supplementary Fig. 11), both of which are consistent with the effects of Smad3 targeting by three different shRNAs.

In contrast, inhibition of myostatin by follistatin resulted in multitudes of very small nascent myofibres, and proliferating $Myf\text{-}5^+$ myogenic cells that persisted within the injured area (Supplementary Fig. 11). Such a defect in myogenic differentiation was also evident when both TGF-β and myostatin were neutralized; strongly suggesting that myostatin, which remains constant with age (Fig. 1), is required for productive differentiation of myogenic cells into myofibres. Distinct regenerative outcomes, where neutralization of TGF-β but not myostatin restores productive repair to old muscle, confirms that TGF-β is the main age-specific local inhibitor in skeletal muscle niche.

Comprehensively, this work establishes that productive muscle repair is determined by an antagonistic balance between the levels of TGF-β/pSmad3 (low in young and high in old) and the activation of Notch (high in young and low in old)[3], which regulates the levels of four distinct CDK inhibitors in resident stem cells. An age-specific shift in either pathway would suffice for the upregulation of CDK inhibitors and diminished proliferation of muscle stem cells. Deregulation of both Notch and TGF-β/pSmad3 thus assures the lack of satellite-cell activation in the context of aged microniche.

Our data indicate that TGF-β does not directly inhibit the expression of Notch ligand Delta, or the levels of active Notch in satellite cells (Supplementary Fig. 12). Likewise, activation of Notch does not directly diminish the TGF-β/pSmad3 levels (Figs 2 and 3). Hence, the

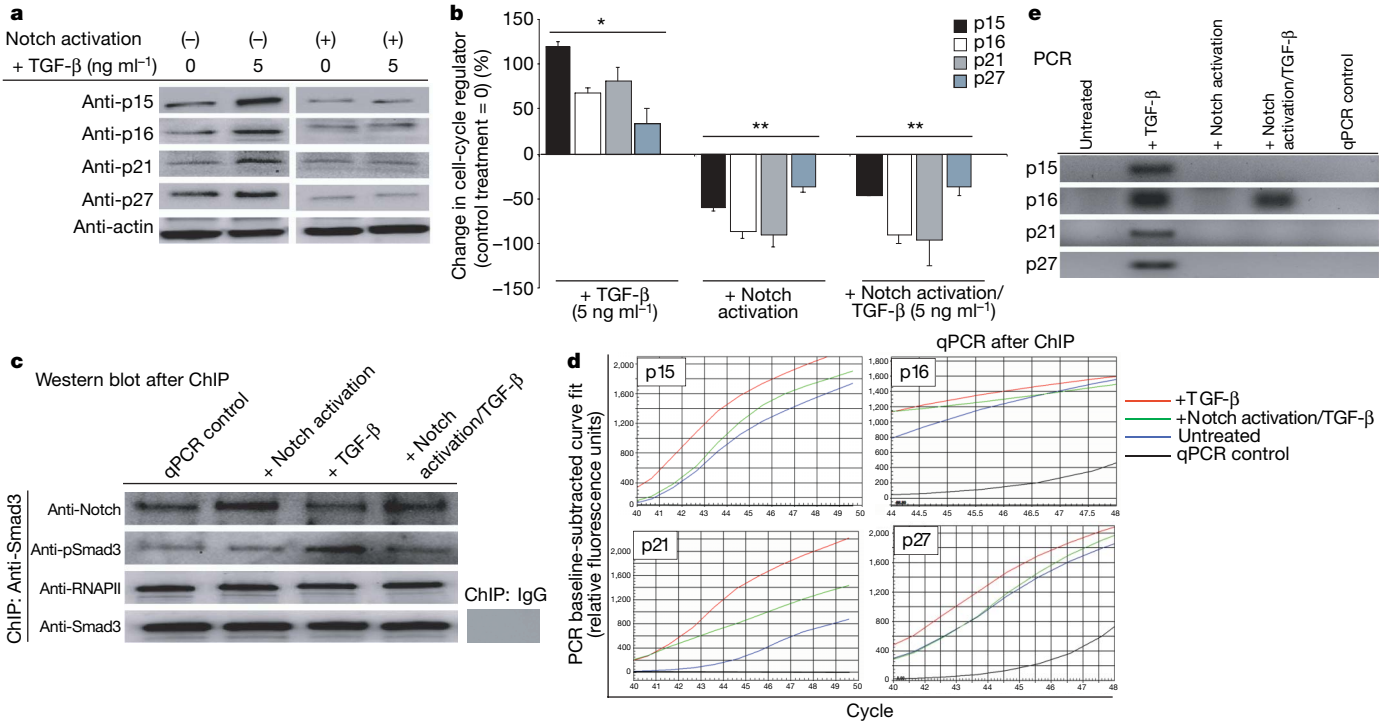


Figure 2 | Notch removes pSmad3 from the 5′ regulatory regions of CDK inhibitors. **a**, Satellite cells treated with TGF-β1, with or without Notch activation, were analysed by western blotting for p15, p16, p21 and p27; quantified in **b**. $*P \leq 0.05$ compared with untreated control (0); $**P \leq 0.05$ compared with TGF-β. Smad3 co-precipitated proteins (**c**) were resolved by western blot; genomic DNA (**d**) was analysed by RT–qPCR, using primers specific for 5′ regions of CDK inhibitors; data are means ± s.d., $n = 3$. **e**, qPCR reactions after 44 cycles revealed fragments of expected molecular weight on agarose gel.

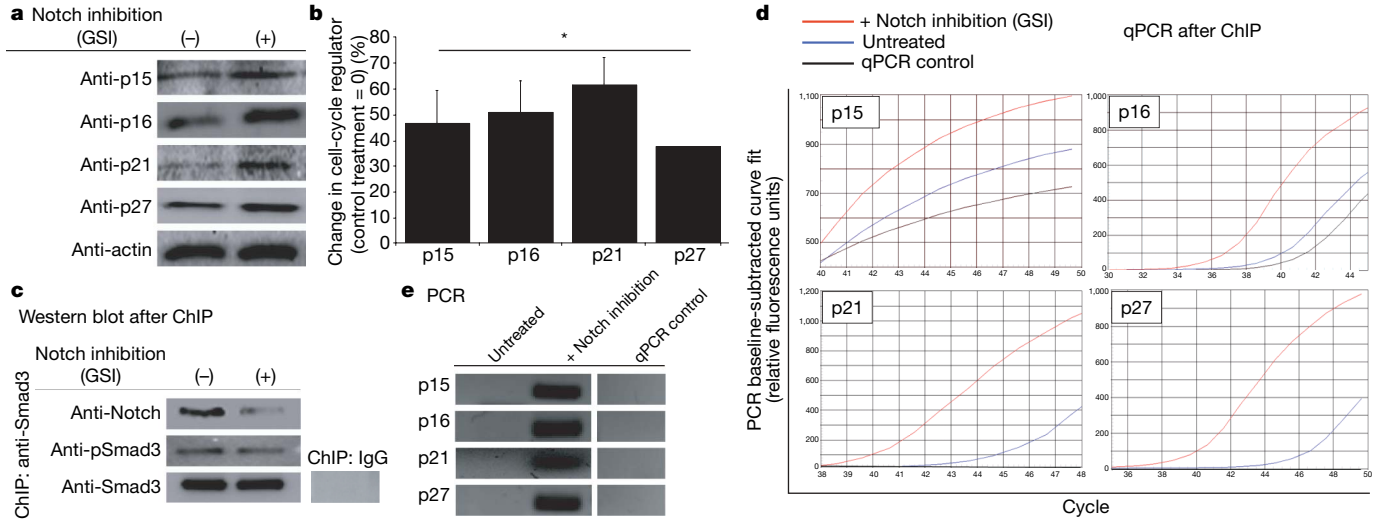


Figure 3 | Inhibition of endogenous Notch upregulates CDK inhibitors. **a**, Compared with untreated cells (−), Notch inhibition by GSI caused prompt upregulation of p15, p16, p21 and p27; western blot assays quantified in **b**; $*P \leq 0.05$ compared with untreated control. In ChIP assay, Smad3 co-precipitated proteins were resolved by western blot (**c**), genomic DNA (**d**) was analysed by RT–qPCR, using primers to p15, p16, p21 and p27 5′ regulatory regions; data are means ± s.d., $n = 3$. **e**, qPCR reactions after 44 cycles revealed fragments of expected molecular weight on agarose gel.

age-specific changes in reciprocal activation of Notch and TGF-β/pSmad3 seem to initiate independently of each other, and future work will identify molecular triggers causing such imbalance. Additional studies will also answer whether excessive TGF-β found in old muscle ECM reflects only local secretion or also results from higher levels of TGF-β in the aged circulation, and whether local and systemic TGF-β are connected and provide feedbacks for each other. Importantly, this work has established that one factor in the ageing of

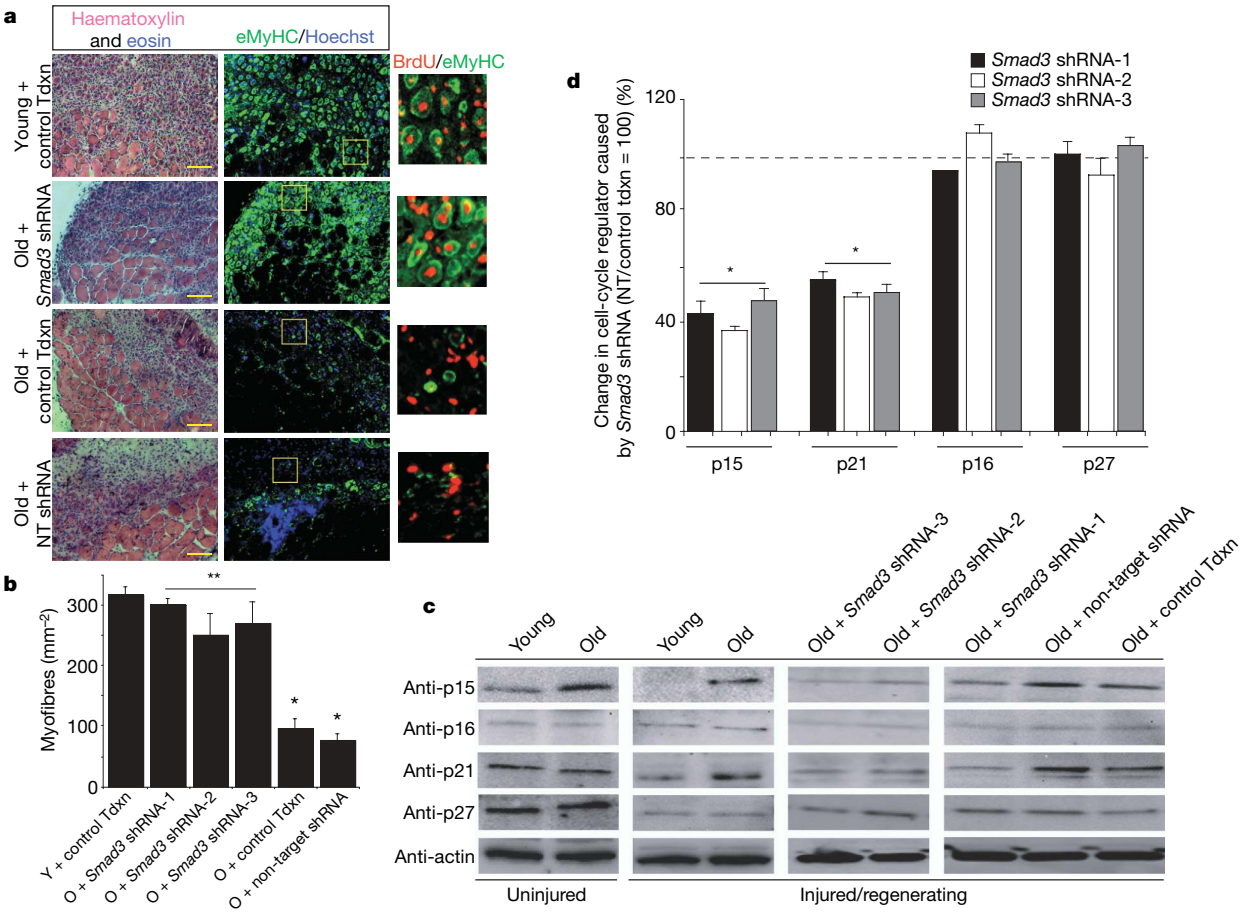


Figure 4 | *Smad3* shRNA rescues responses of satellite cells in old niche *in vivo*. Lentiviral-infected muscle (control transduction (Tdxn), *Smad3* shRNAs, non-target shRNA) was analysed 5 days after injury. **a**, Haematoxylin and eosin 10-μm cryosections of *Smad3* shRNA-1; quantified in **b**; $n = 15$, *, **$P \leq 0.05$; individual *Smad3* shRNA-2, -3 data in Supplementary Fig. 10a; immunodetection of eMyHC myofibres with $BrdU^+$ nuclei. Scale bar, 100 μm; O, old; Y, young. **c**, Isolated satellite cells analysed by western blot for p15, p16, p21 and p27; quantified in **d**; $*P \leq 0.05$, compared with non-target shRNA, control tdxn; data are means ± s.d., $n = 3$.

skeletal muscle—and perhaps other organ niches—seems to be a self-imposed inhibition of regenerative potential.

METHODS SUMMARY

Young (~2 month) and old (~24 month) C57 BL/6 mice were from Jackson Laboratories and the National Institute on Ageing. Skeletal muscle was cardiotoxin injured[3]. Co-injection with anti-TGF-β neutralizing antibody, TGF-β, follistatin or IgG was performed in some experiments. Muscle was isolated 1–5 days after injury and prepared for cryosectioning/histology, western blotting or *in vitro* culturing of myofibre explants and satellite cells.

Immunocytochemistry/histological analyses were performed as described[3,11]. Samples were analysed with a Zeiss Axio Imager A1, imaged with an Axiocam MRc camera and AxioVision software. Western blots were analysed with Bio-Rad Gel Doc/Chemi Doc Imaging System and Quantity One software.

Transwell co-cultures (1 μm) were performed with isolated satellite cells and young/old myofibres. Conditioned supernatants were analysed for secreted bioactive TGF-β1 levels in enzyme-linked immunosorbent assay (ELISA)-based cytokine antibody arrays (Raybiotech). CDK inhibitors were induced *in vitro* by culturing isolated satellite cells with TGF-β1 or myostatin. Notch was activated by immobilized Delta or EDTA exposure before seeding.

Lentiviral transduction was performed with distinct *Smad3* shRNAs: accession number NM_016769.2. (shRNA-2): CCGGCCTTACCACTATCAGAGA GTACTCGAGTACTCTCTGATAGTGGTAAGGTTTTTG, (shRNA-3)-CCGG CTGTCCAATGTCAACCGGAATCTCGAGATTCCGGTTGACATTGGACAG-TTTTTG. *Smad3* shRNA cocktail (shRNA-1, see Methods) yielded results similar to shRNAs 2, 3. Control transduction with non-target shRNA or GFP transduction lentiviral particles was also performed.

BrdU was injected intraperitoneally, 3 days after injury. Tissues were analysed 5 days after injury for regenerative responses and transduction levels by histological analysis, western blotting and PCR with reverse transcription (RT–PCR) (Bio-Rad iQ5).

In ChIP assays (Upstate), satellite cells underwent treatments for 24 h, followed by DNA shearing and immunoprecipitations. Proteins co-precipitated with pSmad3 were analysed by western blot. pSmad3 co-precipitated DNA was analysed using primers to the gene regulatory regions of p15, p16, p21 and p27. Real-time qPCR samples were analysed using a Bio-Rad iQ5 real-time PCR detection system, with iQ5 optical system software. Standard PCR samples were analysed with a Bio-Rad iQ5 thermal cycler.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 6 November 2007; accepted 28 April 2008.
Published online 15 June 2008.

1. Grounds, M. D. Age-associated changes in the response of skeletal muscle cells to exercise and regeneration. *Ann. NY Acad. Sci.* **854,** 78–91 (1998).
2. Renault, V., Thornell, L. E., Eriksson, P. O., Butler-Browne, G. & Mouly, V. Regenerative potential of human skeletal muscle during aging. *Aging Cell* **1,** 132–139 (2002).
3. Conboy, I. M., Conboy, M. J., Smythe, G. M. & Rando, T. A. Notch-mediated restoration of regenerative potential to aged muscle. *Science* **302,** 1575–1577 (2003).
4. Wagers, A. J. & Conboy, I. M. Cellular and molecular signatures of muscle regeneration: current concepts and controversies in adult myogenesis. *Cell* **122,** 659–667 (2005).
5. Collins, C. A. & Partridge, T. A. Self-renewal of the adult skeletal muscle satellite cell. *Cell Cycle* **4,** 1338–1341 (2005).
6. Morgan, J. E. *et al.* Myogenic cell proliferation and generation of a reversible tumorigenic phenotype are triggered by preirradiation of the recipient site. *J. Cell Biol.* **157,** 693–702 (2002).
7. Schultz, E. & Lipton, B. H. Skeletal muscle satellite cells: changes in proliferation potential as a function of age. *Mech. Ageing Dev.* **20,** 377–383 (1982).
8. Brack, A. S. *et al.* Increased Wnt signaling during aging alters muscle stem cell fate and increases fibrosis. *Science* **317,** 807–810 (2007).
9. Conboy, I. M. *et al.* Rejuvenation of aged progenitor cells by exposure to a young systemic environment. *Nature* **433,** 760–764 (2005).
10. Li, L. & Xie, T. Stem cell niche: structure and function. *Annu. Rev. Cell Dev. Biol.* **21,** 605–631 (2005).
11. Carlson, M. E. & Conboy, I. M. Loss of stem cell regenerative capacity within aged niches. *Aging Cell,* (2007).
12. Massague, J. TGF-beta signal transduction. *Annu. Rev. Biochem.* **67,** 753–791 (1998).
13. Massague, J. & Chen, Y. G. Controlling TGF-beta signaling. *Genes Dev.* **14,** 627–644 (2000).
14. Derynck, R. & Zhang, Y. E. Smad-dependent and Smad-independent pathways in TGF-beta family signalling. *Nature* **425,** 577–584 (2003).
15. Feng, X. H. & Derynck, R. Specificity and versatility in tgf-beta signaling through Smads. *Annu. Rev. Cell Dev. Biol.* **21,** 659–693 (2005).
16. Natarajan, E. *et al.* A keratinocyte hypermotility/growth-arrest response involving laminin 5 and p16INK4A activated in wound healing and senescence. *Am. J. Pathol.* **168,** 1821–1837 (2006).
17. Ito, T., Sawada, R., Fujiwara, Y., Seyama, Y. & Tsuchiya, T. FGF-2 suppresses cellular senescence of human mesenchymal stem cells by down-regulation of TGF-beta2. *Biochem. Biophys. Res. Commun.* **359,** 108–114 (2007).
18. Untergasser, G. *et al.* Profiling molecular targets of TGF-beta1 in prostate fibroblast-to-myofibroblast transdifferentiation. *Mech. Ageing Dev.* **126,** 59–69 (2005).
19. Olson, N. E., Kozlowski, J. & Reidy, M. A. Proliferation of intimal smooth muscle cells. Attenuation of basic fibroblast growth factor 2-stimulated proliferation is associated with increased expression of cell cycle inhibitors. *J. Biol. Chem.* **275,** 11270–11277 (2000).
20. Joulia, D. *et al.* Mechanisms involved in the inhibition of myoblast proliferation and differentiation by myostatin. *Exp. Cell Res.* **286,** 263–275 (2003).
21. Lin, J. *et al.* P27 knockout mice: reduced myostatin in muscle and altered adipogenesis. *Biochem. Biophys. Res. Commun.* **300,** 938–942 (2003).
22. Rao, P. & Kadesch, T. The intracellular form of notch blocks transforming growth factor beta-mediated growth arrest in Mv1Lu epithelial cells. *Mol. Cell. Biol.* **23,** 6694–6701 (2003).
23. Janzen, V. *et al.* Stem-cell ageing modified by the cyclin-dependent kinase inhibitor p16INK4a. *Nature* **443,** 421–426 (2006).
24. Molofsky, A. V. *et al.* Increasing p16INK4a expression decreases forebrain progenitors and neurogenesis during ageing. *Nature* **443,** 448–452 (2006).
25. Krishnamurthy, J. *et al.* p16INK4a induces an age-dependent decline in islet regenerative potential. *Nature* **443,** 453–457 (2006).
26. Stepanova, L. & Sorrentino, B. P. A limited role for p16Ink4a and p19Arf in the loss of hematopoietic stem cells during proliferative stress. *Blood* **106,** 827–832 (2005).
27. Husmann, I., Soulet, L., Gautron, J., Martelly, I. & Barritault, D. Growth factors in skeletal muscle regeneration. *Cytokine Growth Factor Rev.* **7,** 249–258 (1996).
28. McPherron, A. C., Lawler, A. M. & Lee, S. J. Regulation of skeletal muscle mass in mice by a new TGF-beta superfamily member. *Nature* **387,** 83–90 (1997).
29. Barcellos-Hoff, M. H. Latency and activation in the control of TGF-beta. *J. Mammary Gland Biol. Neoplasia* **1,** 353–363 (1996).
30. Svoboda, P. Off-targeting and other non-specific effects of RNAi experiments in mammalian cells. *Curr. Opin. Mol. Ther.* **3,** 248–257 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank R. Derynck and M. Conboy for discussions. This work was supported by National Institutes of Health (NIH) R01 (AG027252), NIH R21 (AG27892) and Ellison's Medical Foundation grants to I.M.C., and a Pre-doctoral Training Fellowship from the California Institute for Regenerative Medicine training grant to M.E.C.

Author Contributions M.E.C. performed all experiments, analysed the data and contributed to the writing of the manuscript; M.H. performed preliminary experiments for Fig.1a, b, d and Supplementary Figs 4 and 11; and I.M.C. designed the study, participated in experiments, interpreted the data and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to I.M.C. (iconboy@berkeley.edu).

METHODS

Animal strains. Young (2–3 month) and old (22–24 month) C57 BL/6 male mice were obtained from pathogen-free breeding colonies at Jackson Laboratories and the National Institute on Ageing, respectively. Animals were housed at the Northwest Animal Facility, University of California, Berkeley. Animal procedures were conducted in accordance with the Administrative Panel on Laboratory Animal Care at University of California, Berkeley.

Muscle injury. Tibialis anterior and gastrocnemius muscles of mice were injected with 5 ng cardiotoxin (CTX-1) (Sigma), suspended in 1 × PBS at four or five sites in each muscle (10 μl per site), using a 28-gauge needle. In some experiments anti-TGF-β neutralizing antibody, TGF-β or control goat IgG (all at 500 ng ml^{-1} final concentration) were co-injected with CTX-1, following the same protocol as described above for CTX alone. Uninjured or injured/regenerating muscle tissue was isolated 1–5 days after injury. Whole muscle was prepared for cryosectioning and histological analysis, western blotting or for use in culturing of isolated satellite cells *in vitro*.

Muscle isolation and satellite-cell culture. Myofibre explants and satellite cells were generated from C57 BL/6 mice as described previously[3,11]. Briefly, whole muscle underwent enzymatic digestion at 37 °C in DMEM (Invitrogen)/1% penicillin–streptomycin (Invitrogen)/125 U ml^{-1} Collagenase Type IIA (Sigma) solution. Bulk myofibres with associated satellite cells (located beneath the basement membrane and above the plasma membrane) were purified away from muscle interstitial cells, tendons, etc. by multiple rounds of trituration, sedimentation and washing. Satellite cells were isolated from purified myofibres by subsequent enzymatic digestion with Collagenase Type IIA and Dispase (Sigma), followed by sedimentation, washing and fine-mesh straining procedures, as in refs 3 and 11. Purified satellite cells and satellite cell-stripped myofibres were used in subsequent experiments, as described below. The approximate 95% purity of satellite cells was routinely confirmed by generation of proliferating fusion-competent myoblasts after 24 h in growth medium (Ham's F-10 (Mediatech), 20% FBS (Mediatech), 5 ng ml^{-1} FGF (Chemicon) and 1% penicillin–streptomycin); and myotube formation after 48 h in DMEM, +2% horse serum. The efficiency of satellite-cell depletion from myofibres was routinely confirmed by the absence of such myogenic potential. Satellite cells were cultured on ECM:PBS-coated plates (1:500; BD Biosciences).

Transwell co-cultures of myofibres and satellite cells. Transwell co-cultures (1.0 μm, Corning) were used for culturing isolated satellite cells with young or old myofibres. Satellite cells were seeded onto ECM-coated plates in OPTI-MEM (Invitrogen), +5% FBS. Transwell inserts, containing isolated myofibre explants from young or old muscle, were placed over satellite cells and cultured for 72–96 h before lysing for western blot analysis (see below). Supernatants conditioned for 24 h, from both young and old control myofibre explants, were used to analyse secreted bioactive TGF-β1 levels (immunodetection of active protein) in ELISA-based cytokine antibody arrays (RayBiotech).

Western blot analysis. Myofibre and satellite-cell lysates were prepared in lysis buffer (50 mM Tris, 150 mM NaCl, 1% NP40, 0.25% sodium deoxycholate and 1 mM EDTA, pH 7.4). Protease inhibitor cocktail (Sigma) and 1 mM PMSF were added before use. Phosphatase activity was inhibited by 1 mM sodium fluoride and 1 mM sodium orthovanadate for pSmad immunodetection. Approximately 30 μg protein extract were run on pre-cast SDS PAGE gels (Biorad). Primary antibodies were diluted in 5% non-fat milk/1 × PBST, and nitrocellulose membranes were incubated with antibody mixtures overnight at 4 °C. HRP-conjugated secondary antibodies (Santa Cruz Biotech) were diluted 1:1,000 in 1 × PBST/1% BSA and incubated for 1 h at room temperature. Blots were developed using Western Lightning ECL reagent (Perkin Elmer), and analysed with Bio-Rad Gel Doc/Chemi Doc Imaging System and Quantity One software. Results of multiple assays were quantified by digitizing the data and normalizing pixel density of examined protein by actin-specific pixel density.

Immunocytochemistry and histological analysis. Muscle tissue was treated in a 25% sucrose/PBS solution, frozen in OCT compound (Tissue Tek), cryosectioned at 10 μm (Thermo Shandon Cryotome E) and immunostained as previously described[3,11]. Immunostaining or haematoxylin and eosin staining were performed on cryosections. For indirect immunoflourescence, sections were permeabilized in (PBS, +1% FBS, +0.25% Triton X-100), incubated with primary antibodies for 1 h at room temperature in PBS, +1%FBS, and then with fluorophore-conjugated, species-specific secondary antibodies for 1 h at room temperature (1:500 in PBS, +1%FBS). pSmad3- and BrdU-specific immunostaining required additional nuclear permeabilization and DNA-chromatin denaturation with 4 N HCL. Nuclei were visualized by Hoechst staining for all immunostains. Biologically active TGF-β and myostatin proteins (immunodetectible ligands cleaved from the latent complex)[14,29] were also examined. Samples were analysed at room temperature with a Zeiss Axio Imager A1, and imaged with an Axiocam MRc camera and AxioVision software.

***In vitro* induction of cell-cycle regulators by TGF-β and myostatin.** Activated-by-injury satellite cells were isolated from injured muscle. Cells were seeded onto ECM-coated, six-well plates at a uniform density of 1.2×10^5 cells per well, and cultured for 24 h in OPTI-MEM +1% mouse sera, containing various levels of TGF-β1 or myostatin. For Notch activation, ECM-coated plates were coated with 2 μg ml^{-1} Delta-4 (DLL4) overnight at 37 °C, before seeding satellite cells. Alternatively, Notch activation was induced by exposing satellite cells to 5 mM EDTA for 15 min at 37 °C before seeding. Cells lysates were prepared and used for western blotting analysis (described above).

shRNA delivery by lentiviral transduction. Young and old tibialis anterior and gastrocnemius muscles were infected, *in vivo*, with control non-target shRNA (Sigma SHC002V), GFP (Sigma SHC003V) or pLKO.1-puro (Sigma SHC001V) control transduction lentiviral particles (at least 10^6 transducing units per millilitre, as determined by p24 antigen ELISA titre). Mouse *Smad3* shRNA-producing lentiviral particles (also obtained from Sigma) were used for *in vivo* transduction experiments (target-set generated from accession number NM_016769.2: (1) CCGGCCCATGTTTCTGCATGGATTTCTCGAGAAATCC ATGCAGAAACATGGGTTTTTG; (2) CCGGCCTTACCACTATCAGAGAGTA CTCGAGTACTCTCTGATAGTGGTAAGGTTTTTG; (3) CCGGCTGTCCAA TGTCAACCGGAATCTCGAGATTCCGGTTGACATTGGACAGTTTTTG; (4) CCGGGCACACAATAACTTGGACCTACTCGAGTAGGTCCAAGTTATTGT-GTGCTTTTTG; (5) CCGGCATCCGTATGAGCTTCGTCAACTCGAGTTGAC GAAGCTCATACGGATGTTTTTG). shRNAs were used in a *Smad3* shRNA cocktail (shRNAs 1–5), designated *Smad3* shRNA-1, or individually (shRNA(2) designated *Smad3* shRNA-2 and shRNA(3) designated *Smad3* shRNA-3). Skeletal muscle was infected by intramuscular injection of lentiviral particles (about 50,000 TU) with a 28-gauge needle on multiple consecutive days, before or coincident with CTX-1 injury. To examine cell proliferation, 50 μl of 10 mM BrdU was injected intraperitoneally at 3 days after injury. Tissues were analysed for regenerative responses and transduction levels at 5 days after injury by cryo-sectioning of whole tissues, or western blotting of satellite-cell lysates (described above). Transcript levels were analysed using SuperScript RT–PCR kit (Invitrogen) for amplification of Smad3 (F = CTGGCTACCTGAGTGAAGATGGAGA, R = AAAGACCTCCCCTCCG ATGTAGTAG) and GAPDH (F = TGAGGCCGGTGCTGAGTATGTCGTG, R = TCCTTGGAGGCCATGTAGGCCAT). Amplification products (25–40 cycles on BioRad iQ5) were examined and confirmed for predicted molecular masses on EtBr-stained 2% agarose gels.

ChIP assays and RT–qPCR/PCR. Isolated satellite cells were treated with TGF-β1 only, activation of Notch only, TGF-β1/Notch together, Notch inhibition (GSI) or untreated (as described above). After culture for 24 h, satellite cells were fixed with 1% PFA and ChIP assay was performed according to manufacturer's guidelines (Upstate). Fragments of about 500 base pairs were produced by shearing DNA with attached proteins (confirmed by EtBr-stained gels), and precipitated with antibodies to DNA-bound protein. Proteins that co-precipitated with pSmad3 were analysed by western blot, using indicated antibodies. DNA that co-precipitated with pSmad3 was analysed using primers specific for the 5′ gene regulatory regions of p15, p16, p21 and p27 (p15 F = TCACCGAAGCTACTGGGTCT , R = GTTCAGGGCGTTGGGATCT; p16 F = GTCACACGACTGGGCGATT, R = GTTGCCCATCATCATCACCT and F = GATGACTTCACCCCGTCACT, R = AACACCCCTGAAAACACTGC/GT CCCTCCTTCCTTCCTCTG; p21 F = CCGCGGTGTCAGAGTCTA, R = CAT GAGCGCATCGCAATC; p27 F = AGCCTACGCTCCGACTGTT, R = AGTT CTGCGACTGCACACAG and F = CTAGCCACCGAAGCTCCTAA, R = AGT TCTGCGACTGCACACAG and F = CTGGCTCTGCTCCATTTGAC, R = G GCTCCCGTTAGACACTCTC). GAPDH primers (see above) were used as control for Smad3 non-enriched genomic regions. Primers were designed with OligoPerfect Designer (Invitrogen). For RT–qPCR, samples were analysed using a Bio-Rad iQ5 real-time PCR detection system, with iQ5 optical system software. For PCR, samples were amplified with Platinum Taq DNA Polymerase (Invitrogen) and analysed on a Bio-Rad iQ5 system. After 40–55 cycles of amplification, fragments produced from each primer set were examined and confirmed for their predicted molecular masses on EtBr-stained 2% agarose gels. Fifty-five cycles of amplification were used for negative control PCR reactions.

Reagents. Antibodies to BrdU (ab6326), activated Notch1 (ab8925) and CHiP grade Smad3 (ab287379) were purchased from Abcam. Antibody to developmental eMyHC (clone RNMy2/9D2) was acquired from Vector Laboratories. Antibodies to desmin (clone DE-U-10, Cat#D8281), laminin (L9393), actin (A5060) and follistatin (F2177) were acquired from Sigma. Bioactivity-neutralizing antibodies against TGF-β1/2/3 were obtained from R&D Systems (MAB1835) and Santa Cruz Biotechnologies (sc7892). Antibodies to myostatin (sc-34781), follistatin (sc-30194), TGF-β1 (sc146), phosphorylated-smad3 (sc11769), smad6 (sc13048), smad7 (sc11392), p15 (sc613), p16 (sc1207 and sc1661), p21 (sc756), p27 (sc776), RNAP II (sc899),

Myf5 (sc31946) and goat/rabbit IgG were acquired from Santa Cruz Biotechnologies. Smad3 antibody (06-920) was obtained from Upstate. Fluorophore-conjugated secondary antibodies (Alexa Fluor) were supplied by Invitrogen. HRP-conjugated secondary antibodies were purchased from Santa Cruz Biotechnologies. BrdU labelling reagent was obtained from Sigma. Recombinant human TGF-β1 (RD 240B), recombinant mouse DLL4 (RD 1389) and recombinant mouse myostatin (788-G8-010) were obtained from R&D Systems. TGF-β RI Kinase Inhibitor (50 nM) (#616451) and gamma-Secretase Inhibitor X (50 nM) (#565771) were purchased from Calbiochem. Two × SYBR-Green RT–PCR reaction mixture was purchased from Bio-Rad, and SuperScript One-Step RT–PCR (#10928) from Invitrogen.

Statistical analysis. Quantified data are expressed as mean ± s.d. Significance testing was performed using one-way analysis of variance, with an alpha level of 0.01–0.05, to compare data from different experimental groups. A minimum of three replicates were performed for each described experimental condition.

LETTERS

Switch of *rhodopsin* expression in terminally differentiated *Drosophila* sensory neurons

Simon G. Sprecher[1] & Claude Desplan[1]

Specificity of sensory neurons requires restricted expression of one sensory receptor gene and the exclusion of all others within a given cell. In the *Drosophila* retina, functional identity of photoreceptors depends on light-sensitive Rhodopsins (Rhs). The much simpler larval eye (Bolwig organ) is composed of about 12 photoreceptors, eight of which are green-sensitive (Rh6) and four blue-sensitive (Rh5)[1]. The larval eye becomes the adult extraretinal 'eyelet' composed of four green-sensitive (Rh6) photoreceptors[2,3]. Here we show that, during metamorphosis, all Rh6 photoreceptors die, whereas the Rh5 photoreceptors switch fate by turning off Rh5 and then turning on Rh6 expression. This switch occurs without apparent changes in the programme of transcription factors that specify larval photoreceptor subtypes. We also show that the transcription factor Senseless (Sens) mediates the very different cellular behaviours of Rh5 and Rh6 photoreceptors. Sens is restricted to Rh5 photoreceptors and must be excluded from Rh6 photoreceptors to allow them to die at metamorphosis. Finally, we show that Ecdysone receptor (EcR) functions autonomously both for the death of larval Rh6 photoreceptors and for the sensory switch of Rh5 photoreceptors to express Rh6. This fate switch of functioning, terminally differentiated neurons provides a novel, unexpected example of hard-wired sensory plasticity.

The adult *Drosophila* eyelet comprises approximately four photoreceptors located between the retina and the optic ganglia[2]. It directly contacts the pacemaker neurons of the adult fly, the lateral neurons[4]. In conjunction with the compound eye and the clock-neuron intrinsic blue-sensitive receptor cryptochrome[3] it helps shift the phase of the molecular clock in response to light. All eyelet photoreceptors express green-sensitive Rh6, and are derived from photoreceptors of the larval eye[2,5,6] that mediate light avoidance and entrainment of the molecular clock by innervating the larval lateral neurons[7–9].

Larval photoreceptors develop in a two-step process during embryogenesis[1,10]. Primary precursors are specified first and develop as the four Rh5-subtype photoreceptors. They signal through Epidermal growth factor receptor (EGFR) to the surrounding tissue to develop as secondary precursors, which develop into the eight Rh6-subtype photoreceptors[1]. Two transcription factors specify larval photoreceptor subtypes[1]. Spalt (Sal) is exclusively expressed in Rh5 photoreceptors, where it is required for Rh5 expression. Seven-up (Svp) is restricted to Rh6 photoreceptors, where it represses *sal* and promotes Rh6 expression. A third transcription factor, Orthodenticle (Otd), expressed in all larval photoreceptors, acts only in the Rh5 subtype to promote Rh5 expression and to repress Rh6 (refs 1, 11).

To address the relation between the larval Rh5 and Rh6 photoreceptors and the adult eyelet, we tracked them through metamorphosis (Fig. 1c–e). To permanently label them, we used UAS-Histone2B::YFP, which is stably incorporated in the chromatin, and thus remains detectable in post-mitotic neurons throughout pupation[12]. Surprisingly, all Rh6 photoreceptors degenerate and disappear during early phases of metamorphosis (Fig. 1d). In contrast, Rh5 photoreceptors can be followed throughout pupation (Fig. 1g–i). Expression of Rh5 ceases during early stages of pupation and, at mid-pupation, neither Rh5 nor Rh6 can be detected (Fig. 1h). About four cells are still present, however, and can be identified by *rh5*-Gal4/UAS-*H2B::YFP* (Fig. 1h) or *GMR*-Gal4/UAS-*H2B::YFP* (data not shown). Eyelet photoreceptors only express Rh6, even though H2B::YFP driven by *rh5*-Gal4 is detectable in those cells (Fig. 1i). Therefore, the four larval Rh5 photoreceptors must switch *rhodopsin* expression at metamorphosis to give rise to the four eyelet Rh6 photoreceptors (Fig. 1j). The remaining eight Rh6 photoreceptors die (Fig. 1f), their axon becoming fragmented before disappearing (Fig. 1m, n). A 'memory experiment' (*rh5*-Gal4/UAS-Flp;*Act*-FRT > STOP > FRT-*nlacZ*) also showed that eyelet Rh6 photoreceptors did express Rh5 earlier (Fig. 1k, l).

We further verified the death of Rh6 photoreceptors and transformation of Rh5 photoreceptors by three independent sets of experiments.

First, we ablated Rh5 photoreceptors by expressing pro-apoptotic genes *rpr* and *hid* (*rh5*-Gal4/UAS-*rpr*,UAS-*hid*). This results in the absence of larval Rh5 photoreceptors and the complete absence of the eyelet (Fig. 2g, h)[4]. Conversely, preventing cell death of the Rh6 subtype by expressing the apoptosis inhibitor *p35* (*rh6*-Gal4/UAS-*p35*) leads to an eyelet that consists of 12 photoreceptors, all expressing Rh6 (Fig. 2i).

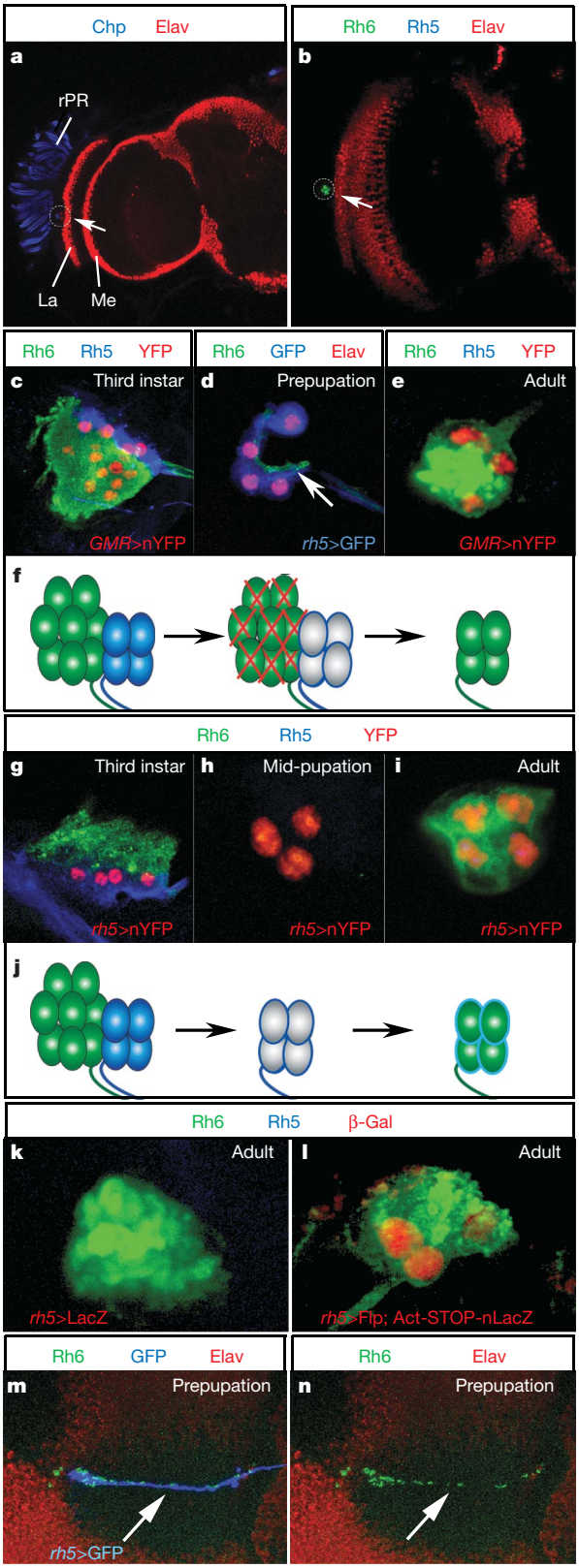
Second, we blocked development of larval Rh6 photoreceptors by expressing a dominant negative form of EGFR (*so*-Gal4/UAS-*H2B::YFP;* UAS-$EGFR^{DN}$) (Fig. 2a)[1]. The eyelet of these animals is not affected and three or four cells express Rh6 normally (Fig. 2d). This shows that larval Rh6 photoreceptors do not contribute to the eyelet.

Third, we analysed the expression of Sal (Rh5-subtype specific) and Svp (Rh6-subtype specific) in the adult eyelet: eyelet photoreceptors still express Sal, but not Svp even though these photoreceptors now express Rh6 (Fig. 2b, c, e, f). Rh5 requires Sal expression in the Bolwig organ, but Otd function is also necessary to activate Rh5 and to repress Rh6. In *otd* mutants, larval Rh5 photoreceptors marked by Sal express Rh6 and lack Rh5 expression, thus mimicking the switch at metamorphosis[1]. Thus, Rh6 could be expressed in Rh5 photoreceptors if *otd* function were lost in the eyelet. However, Otd expression does not change during the transition from the Bolwig organ to eyelet (data not shown) although it might be inactive in the eyelet.

What is the trigger that controls the switch from *rh5* to *rh6*? Ecdysone controls many developmental processes during metamorphosis. EcR is expressed during the third larval instar and pupation in all larval photoreceptors and surrounding tissues (Fig. 3a, b, d,

[1]Center for Developmental Genetics, Department of Biology, New York University, 1090 Silver Center, 100 Washington Square East, New York, New York 10003-6688, USA.

e)[13]. To evaluate EcR activity, we used a reporter line in which *lacZ* is under the control of multimerized ecdysone response elements (7×EcRE-*lacZ*)[14]. The expression of *lacZ* is absent until late third instar and prepupation, whereas thereafter all larval photoreceptors (and surrounding tissue) express 7×EcRE-*lacZ* (Fig. 3c, f). EcR expression decreases during late pupation and is no longer detectable by the time Rh6 expression starts in the eyelet (Supplementary Figure 2a, b).



To test the role of ecdysone, we expressed a dominant negative form of EcR specifically in larval Rh5 photoreceptors, while permanently labelling these cells (*rh5*-Gal4/UAS-*H2B*::*YFP*;UAS-EcR^{DN}). This causes no disruption of larval photoreceptor fate, but the eyelet of these animals now consists of four photoreceptors that all express Rh5 instead of Rh6 (Fig. 4B, G). A comparable phenotype is observed after expression of an RNA interference (RNAi) construct for EcR (*rh5*-Gal4/UAS-*H2B*::*YFP*;UAS-EcR^{RNAi})(Fig. 4C, G). Therefore, loss of EcR function prevents larval photoreceptors from switching to Rh6 expression. In both cases, larval Rh6 photoreceptors still degenerate and are not observed in the eyelet (Fig. 4G).

We also expressed the dominant negative form of EcR in Rh6 photoreceptors (*rh6*-Gal4/UAS-*H2B*::*YFP*; UAS-EcR^{DN}). In this case, the Bolwig organ is not affected but the resulting adult eyelet consists of about 12 photoreceptors, all expressing Rh6 (Fig. 4D, H). This presumably results from Rh6 photoreceptors not undergoing apoptosis whereas larval Rh5 photoreceptors still switch expression to Rh6 in the eyelet (Fig. 4H). Expression of UAS-EcR-RNAi in Rh6 photoreceptors (*rh6*-Gal4/UAS-*H2B*::YFP;UAS-EcR^{RNAi}) leads to the same results (Fig. 4E, H).

Although EcR could directly control the switch of *rhodopsin* expression through binding to the promoters of *rh5* and *rh6*, these promoters contain no potential EcR binding sites[15]. Moreover, as no EcR expression is detectable when Rh6 starts to be expressed, this would make it unlikely for EcR to control directly the switch to Rh6 (Supplementary Fig. 2a, b). Finally, only allowing expression of the dominant negative form of EcR starting at mid-pupation (*GMR*-Gal4/Tub-$Gal80^{ts}$,UAS-EcR^{DN}), after *rh5* is switched off, does not prevent activation of Rh6 in the eyelet (Supplementary Fig. 2c, d). Thus EcR most likely acts in an indirect manner in regulating *rhodopsins*, likely through the activation of transcription factors that bind to *rh5* and *rh6* promoters.

The differential response to ecdysone of Rh6 photoreceptors (which die) and of Rh5 photoreceptors (which switch to Rh6) must be due to intrinsic differences between the two subtypes before EcR signalling. Likely candidates are Sal and Svp. However, late misexpression of Svp in Rh5 photoreceptors (*rh5*-Gal4/UAS-*H2B*::YFP;UAS-*svp*) or of Sal in Rh6 photoreceptors (*rh6*-Gal4/UAS-*H2B*::YFP;UAS-*sal*) neither affects *rhodopsin* expression or cell number in the eyelet nor alter the expression of *rhodopsins* in the Bolwig organ (which is only affected by very early expression of these transcription factors, through *so*-Gal4 (ref. 1)). Thus neither Sal nor Svp are sufficient to alter the response of larval photoreceptors to EcR.

Figure 1 | Transformation of the larval eye into the adult eyelet. **a**, **b**, The eyelet locates between the optic ganglia (anti-Elav, red; La, lamina; Me, medulla) and retina (anti-Chp, blue; rPR, retinal photoreceptors). **b**, The eyelet only expresses Rh6 (green). **c**, Larval photoreceptors express Rh6 (green) or Rh5 (blue), nuclei in red (*GMR* > *H2B::YFP*). **d**, Rh6 photoreceptors (green) degenerate during prepupation (arrow), Rh5 photoreceptors (rh5 > GFP, blue) remain, nuclei in red (anti-Elav). **e**, High magnification of eyelet photoreceptors expressing Rh6 (green) not Rh5 (blue), nuclei in red (*GMR* > *H2B::YFP*). **f**, Transformation of larval photoreceptors: Rh6 photoreceptors degenerate, Rh5 photoreceptors remain at prepupation stages but express Rh6 in the eyelet. **g–i**, Rh5 photoreceptors tracked through metamorphosis using *rh5* > *H2B::YFP* (red), anti-Rh6 (green) and anti-Rh5 (blue): **g**, during third-instar larva; **h**, mid-pupation (neither Rh5 nor Rh6 detectable); **i**, eyelet photoreceptors now expressing Rh6. **j**, By mid-pupation, Rh6 photoreceptors have degenerated whereas Rh5 photoreceptors are now empty, they later switch to express Rh6. **k**, No *rh5* > *lacZ* expression can be detected in the eyelet. **l**, Genetic memory experiment (*rh5*-Gal4/UAS-Flp;*Act*-FRT > STOP > FRT-*nlacZ*): *lacZ* detected in eyelet (anti β-Gal, red; anti-RH6, green). **m**, **n**, Projections of Rh6 photoreceptors undergo fragmentation during prepupation (**n**, arrow, labelled with anti-Rh6), whereas Rh5 photoreceptor projections remain (**m**, Rh5 photoreceptor projections are shown by *rh5*-GFP; anti-Elav, red).

An additional factor, independent from *svp* and *sal*, must therefore allow survival of Rh5 photoreceptors, or promote Rh6 photoreceptor death. We found that the transcription factor Sens is specifically expressed in larval Rh5 photoreceptors and remains expressed in all cells in the eyelet (Fig. 2j, k) where it might act to promote cell survival. To test this, we misexpressed *sens* in Rh6 photoreceptors (*rh6*-Gal4/UAS-*H2B*::YFP;UAS-*sens*). This results in an eyelet that consists of 12 photoreceptors, all expressing Rh6 (Fig. 2l). Thus, expression of Sens in Rh6 photoreceptors is sufficient to rescue them from death, without affecting Sal and Svp expression and subtype specification of larval photoreceptors (data not shown).

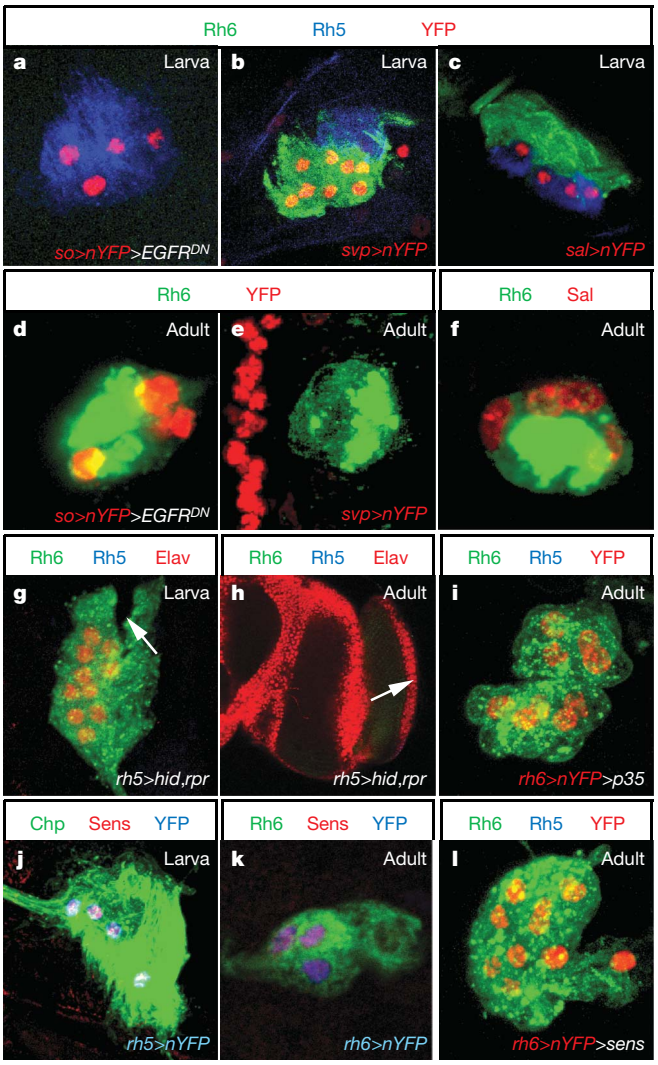


Figure 2 | Larval Rh5 photoreceptors give rise to the eyelet and express Rh5 photoreceptor markers. **a**, Inhibition of larval Rh6 photoreceptor development (*so* > *H2B::YFP,EGFR*^{DN}): only the Rh5 subtype is present in larvae (blue)[1], whereas (**d**) the eyelet remains unaffected (anti-Rh6, green; anti-Rh5, blue; anti-YFP, red). **b**, In the larva, Svp (*svp* > *H2B::YFP*, red) is expressed in Rh6 (green) but not Rh5 photoreceptors (blue), whereas (**e**) the eyelet does not express Svp (anti-Rh6, green). **c**, In the larva, Sal (*sal* > *H2B::YFP*, red) is expressed in Rh5 photoreceptors (blue) but not in Rh6 photoreceptors (green) (**f**). Sal (red) is expressed in the eyelet (anti-Rh6, green). **g**, **h**, *rh5*-Gal4,UAS-*hid,rpr* ablates Rh5 photoreceptors (**g**, arrow) and eyelet (**h**, arrow) (anti-Rh5, blue; anti Rh6, green; anti Elav, red). **i**, *rh6*-Gal4/UAS-*p35* prevents apoptosis of Rh6 photoreceptors (anti-Rh6, green; anti-YFP, red). **j**, **k**, Sens (red) is detected in the larval Rh5 subtype (*rh5* > *H2B::YFP*, blue; anti-Chp, green) and the eyelet (*rh6* > *H2B::YFP*; anti-YFP, blue; anti-Rh6, green). **l**, Misexpression of UAS-*sens* in Rh6 photoreceptors (*rh6* > *H2B::YFP* > *sens*) prevents their apoptosis (anti-Rh6, green; anti-YFP, red; anti-Rh5, blue). The same phenotype is obtained with *GMR*-Gal4;UAS-*sens* (data not shown).

Ecdysone hormonal signalling thus acts in two independent ways during the formation of the adult eyelet. First, it induces the degeneration of the Rh6 subtype, thereby assuring the correct number of eyelet photoreceptors. This apoptotic death requires the absence of Sens, whose expression is restricted to Rh5 photoreceptors that survive. Second, ecdysone signalling is also required to trigger the switch of spectral sensitivity of blue-sensitive (Rh5) larval photoreceptors to green-sensitive (Rh6) eyelet photoreceptors (Fig. 4I).

Thus terminally differentiated sensory neurons switch specificity by turning off one Rhodopsin and replacing it with another. Although examples of such switches in sensory specificity of terminally differentiated, functional, sensory receptors are extremely rare, this strategy might be more common than currently anticipated. In the Pacific pink salmon and rainbow trout, newly hatched fish express an ultraviolet opsin that changes to a blue opsin as the fish ages[16–18]. As in flies, this switch might reflect an adaptation of vision to the changing lifestyle. The maturing salmon, born in shallow water, later migrates deeper in the ocean where ultraviolet does not penetrate. The *rhodopsin* switch in the eyelet may similarly be an adaptation to the deeper location of the eyelet within the head, as light with longer wavelengths (detected by Rh6) penetrates deeper into tissue than light with shorter wavelengths (detected by Rh5).

The eyelet functions with retinal photoreceptors and Cryptochrome to entrain the molecular clock in response to light. The larval eye, on the other hand, functions in two distinct processes: for the entrainment of the clock and for the larva to avoid light[7]. Interestingly, the Rh5 subtype appears to support both functions whereas Rh6 photoreceptors only contribute to clock entrainment (S.G.S., J. Blau and C.D., unpublished observations). Thus, the photoreceptor subtype that supports both functions of the larval eye is the one that is maintained into the adult and becomes the eyelet. Why are Rh6-sensitive photoreceptors not maintained? As these photoreceptors are recruited to the larval eye secondarily, the ancestral Bolwig organ might have had only Rh5 photoreceptors and had to undergo a switch in specificity. Larval Rh5 photoreceptors appear to maintain their overall connectivity to the central pacemaker neurons. However, they are also profoundly restructured and exhibit widely increased connectivity during metamorphosis. This might be due to the increase in number of their target neurons, and the switch of Rh might be part of more extensive plasticity during formation of the eyelet, including increased connectivity and possibly the innervation of novel target neurons.

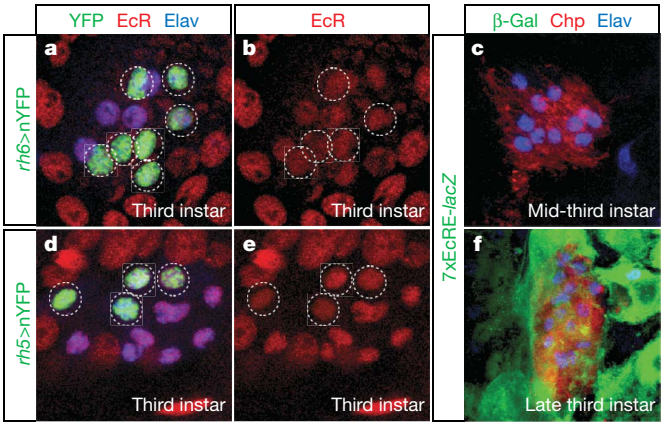


Figure 3 | EcR expression and activity in larval photoreceptors before metamorphosis. **a**, **b**, **d**, **e**, In third-instar larvae, EcR protein (red) is detected in Rh6 photoreceptors (*rh6* > *H2B::YFP*; green) (**a**) and in Rh5 photoreceptors (*rh5* > *H2B::YFP*; green) (**d**); anti-Elav (blue). **b**, **e**, The same as **a**, **d** showing only EcR expression (broken circles highlight Rh6 photoreceptors in **a** and **b**, Rh5 photoreceptors in **d** and **e**). **c**, **f**, EcR activity monitored using 7×EcRE-*lacZ* (β-Gal, green). Anti-Chp (red) and anti-Elav (blue) marked photoreceptors. No EcR activity is detected during mid-third instar (**c**) but is present from late third-instar larval (**f**) to prepupal stages.

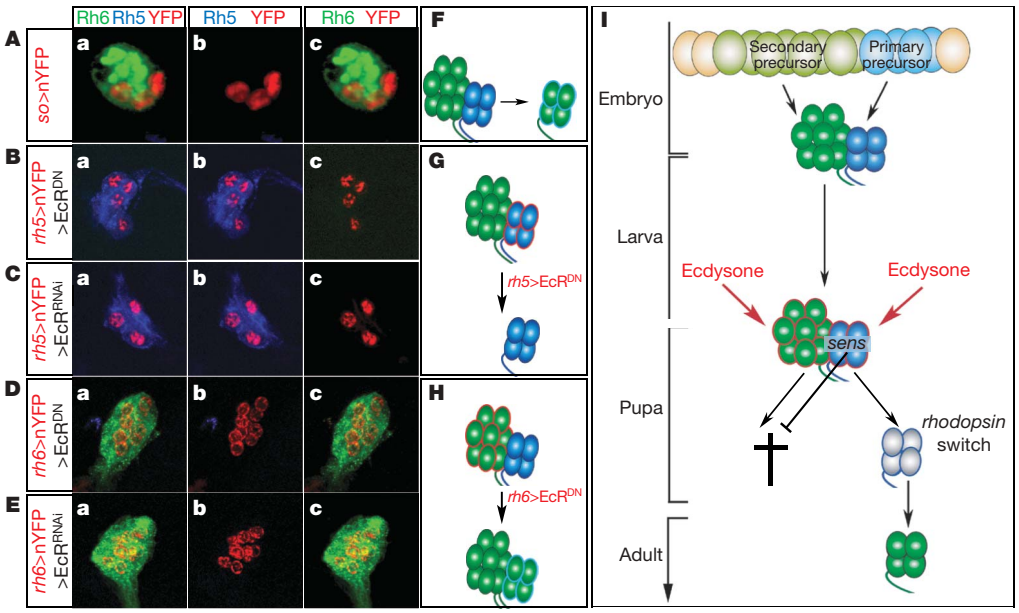


Figure 4 | EcR is required autonomously for the fate switch of Rh5 photoreceptors and apoptosis of Rh6 photoreceptors. Eyelet expressing so-Gal4/UAS-*H2B::YFP* (**Aa**–**Ac**), *rh5*-Gal4,UAS-*H2B::YFP*, and UAS-EcR^{DN} (**Ba**–**Bc**), *rh5*-Gal4,UAS-*H2B::YFP*,UAS-EcR^{RNAi} (**Ca**–**Cc**), *rh6*-Gal4,*UAS*-H2B::YFP,UAS-EcR^{DN} (**Da**–**Dc**) or *rh6*-Gal4,UAS-*H2B::YFP*,UAS-EcR^{RNAi} (**Ea**–**Ec**), anti-Rh6 (green), anti-Rh5 (blue) and anti-YFP (red). Interfering with EcR in Rh5 photoreceptors prevents cells from switching to Rh6 (**b**, **c**). In Rh6 photoreceptors, it prevents apoptosis (**D**, **E**). The eyelet (**F**) after manipulation of Rh5 (**G**) or Rh6 photoreceptors (**H**). **I**, Transformation of the larval eye into the eyelet. EcR function leads to apoptosis of the Rh6 photoreceptors and the switch to Rh6 of Rh5 photoreceptors.

The general model that sensory neurons only express a single sensory receptor gene does not hold true for salmon and the fruitfly[19]. Interestingly, reports from several other species, including amphibians, rodents and humans, show co-expression of opsins[17,18,20–22]. In humans, for instance, it has been proposed that cones first express S opsin and later switch to L/M opsin. However, this likely reflects a developmental process rather than a functional adaptation[22].

We identified two major players in the genetic programme for the transformation of the larval eye to the eyelet. First, EcR acts as a trigger for both *rhodopsin* switch and apoptosis. Surprisingly, the upstream regulators specifying larval photoreceptor-subtype identity, Sal, Svp and Otd, do not contribute to the genetic programme of sensory plasticity of the *rhodopsin* switch. Therefore a novel genetic programme is required for regulating *rhodopsin* expression in the eyelet, which likely depends on downstream effectors of EcR.

Second, larval Rh5 and Rh6 photoreceptors respond differently to ecdysone, either switching *rhodopsin* expression or undergoing apoptosis. This appears to depend on Sens, which is likely to be required for the survival of Rh5 photoreceptors. The role of Sens in inhibiting apoptosis is not unique to this situation: Sens is essential to promote survival of salivary-gland precursors during embryogenesis[23]. The vertebrate homologue of *sens*, *Gfi-1*, acts to inhibit apoptosis of T-cell precursors in haematopoiesis and cochlear hair cells of the inner ear[24,25]. Thus the anti-apoptotic function of Sens/Gfi-1 may be a general property of this molecule.

Ecdysone acts in remodelling neurons during metamorphosis[19,20,22]. In γ-neurons of the mushroom body, a structure involved in learning and memory, ecdysone is required for the pruning of larval processes[26]. Similarly, dendrites of C4da sensory neurons undergo large-scale remodelling that depends on ecdysone signalling[27]. Interestingly, in the moth *Manduca*, 'lateral neurosecretory cells' express cardio-acceleratory peptide 2, which is switched off in response to ecdysone before expression of the neuropeptide bursicon is initiated in the adult[28].

The transformation of larval blue-sensitive photoreceptors to green-sensitive photoreceptors of the eyelet reveals an unexpected example of sensory plasticity by switching *rhodopsin* gene expression in functional, terminally differentiated sensory neurons.

METHODS SUMMARY

***Drosophila* strains and genetics.** The following fly strains were used: yw[122]; *so*-Gal4, *rh5*-Gal4, *rh6*-Gal4, *rh5*-GFP, *rh6*-GFP, *rh5*-lacZ *otd*^{uvi}, *otd*-Gal4 (T. Cook, personal communication), UAS-*sens*[29], *GMR*-Gal4 (Bloomington Drosophila Stock Center), UAS-EcR^{RNAi}, *svp*^{H162}-LacZ, *svp*^{724}-Gal4 (Kyoto Stock Center), *sal*-Gal4, UAS-*EGFR*^{dn}, UAS-*melt*[30], UAS-*EcR*^{DN}, (UAS-*EcR*^{DN} for isoforms A, B1 and B2 all gave comparable results; UAS-*EcR*^{DN}-B2 is used in the figures), 7×EcRE-*lacZ*[14], UAS-*EcR*^{RNAi}, UAS-*H2B::YFP* (anti-GFP antibody/biogenesis recognizes the YFP antigen), UAS-*svp*, UAS-*salm*, UAS-*hid*, UAS-*rpr*, UAS-*GFP*, *Act*-FRT > STOP > FRT-*nlacZ*, UAS-*lacZ*, UAS-*p35*, Tub-Gal80ts (Bloomington).

Immunohistochemistry and preparation of larval and adult specimens. Primary antibodies were rabbit anti-Rh6 1:10,000, mouse anti-Rh5 1:20, rat anti-Elav 1:30 (Developmental Studies Hybridoma Bank (DSHB)), mouse anti-EcR (DSHB, antibodies against EcR-A, EcR-B1 and EcR-B2 gave comparable results; anti-EcR^{common} is used in all figures), sheep anti-GFP (Biogenesis), rabbit anti-Sal 1:200, guinea pig anti-Sens 1:1000, mouse anti-Svp 1:1000, mouse anti-Pros 1:50 (DSHB), mouse anti-Chp 1:10 (DSHB), mouse anti-Pdf 1:30 (DSHB), rat anti-Otd 1:200 and mouse anti-betaGAL 1:20 (DSHB). Secondary antibodies used for confocal microscopic analysis were Alexa-488, Alexa-555 and Alexa-647 generated in goat (Molecular Probes, Invitrogen), all at 1:300–1:500 dilution. Specimens were mounted in Vectashield H-1000 (Vector). Dissection for staining of the larval eye was performed as previously described[1].

Laser confocal microscopy and image processing. A Leica TCS SP laser confocal microscope was used. Optical sections ranged from 0.2 to 2 μm recorded in line average mode with picture size of 512 pixels × 512 pixels, or 1,024 pixels × 1,024 pixels. Captured images from optical sections were arranged and processed using Leica Confocal Software and ImageJ, and imported into Adobe Photoshop.

Received 28 February; accepted 9 May 2008.
Published online 25 June 2008.

1. Sprecher, S. G., Pichaud, F. & Desplan, C. Adult and larval photoreceptors use different mechanisms to specify the same rhodopsin fates. *Genes Dev.* **21**, 2182–2195 (2007).
2. Helfrich-Forster, C. *et al.* The extraretinal eyelet of *Drosophila*: development, ultrastructure, and putative circadian function. *J. Neurosci.* **22**, 9255–9266 (2002).

3. Veleri, S., Rieger, D., Helfrich-Forster, C. & Stanewsky, R. Hofbauer-Buchner eyelet affects circadian photosensitivity and coordinates TIM and PER expression in *Drosophila* clock neurons. *J. Biol. Rhythms* **22,** 29–42 (2007).
4. Mealey-Ferrara, M. L., Montalvo, A. G. & Hall, J. C. Effects of combining a cryptochrome mutation with other visual-system variants on entrainment of locomotor and adult-emergence rhythms in *Drosophila*. *J. Neurogenet.* **17,** 171–221 (2003).
5. Malpel, S., Klarsfeld, A. & Rouyer, F. Larval optic nerve and adult extra-retinal photoreceptors sequentially associate with clock neurons during *Drosophila* brain development. *Development* **129,** 1443–1453 (2002).
6. Yasuyama, K. & Meinertzhagen, I. A. Extraretinal photoreceptors at the compound eye's posterior margin in *Drosophila melanogaster*. *J. Comp. Neurol.* **412,** 193–202 (1999).
7. Mazzoni, E. O., Desplan, C. & Blau, J. Circadian pacemaker neurons transmit and modulate visual information to control a rapid behavioral response. *Neuron* **45,** 293–300 (2005).
8. Busto, M., Iyengar, B. & Campos, A. R. Genetic dissection of behavior: modulation of locomotion by light in the *Drosophila melanogaster* larva requires genetically distinct visual system functions. *J. Neurosci.* **19,** 3337–3344 (1999).
9. Hassan, J., Iyengar, B., Scantlebury, N., Rodriguez Moncalvo, V. & Campos, A. R. Photic input pathways that mediate the *Drosophila* larval response to light and circadian rhythmicity are developmentally related but functionally distinct. *J. Comp. Neurol.* **481,** 266–275 (2005).
10. Daniel, A., Dumstrei, K., Lengyel, J. A. & Hartenstein, V. The control of cell fate in the embryonic visual system by atonal, tailless and EGFR signaling. *Development* **126,** 2945–2954 (1999).
11. Tahayato, A. *et al.* Otd/Crx, a dual regulator for the specification of ommatidia subtypes in the *Drosophila* retina. *Dev. Cell* **5,** 391–402 (2003).
12. Kimura, H. Histone dynamics in living cells revealed by photobleaching. *DNA Repair* **4,** 939–950 (2005).
13. Talbot, W. S., Swyryd, E. A. & Hogness, D. S. *Drosophila* tissues with different metamorphic responses to ecdysone express different ecdysone receptor isoforms. *Cell* **73,** 1323–1337 (1993).
14. Kozlova, T. & Thummel, C. S. Essential roles for ecdysone signaling during *Drosophila* mid-embryonic development. *Science* **301,** 1911–1914 (2003).
15. Devarakonda, S., Harp, J. M., Kim, Y., Ozyhar, A. & Rastinejad, F. Structure of the heterodimeric ecdysone receptor DNA-binding complex. *EMBO J.* **22,** 5827–5840 (2003).
16. Cheng, C. L. & Flamarique, I. N. Chromatic organization of cone photoreceptors in the retina of rainbow trout: single cones irreversibly switch from UV (SWS1) to blue (SWS2) light sensitive opsin during natural development. *J. Exp. Biol.* **210,** 4123–4135 (2007).
17. Cheng, C. L., Flamarique, I. N., Harosi, F. I., Rickers-Haunerland, J. & Haunerland, N. H. Photoreceptor layer of salmonid fishes: transformation and loss of single cones in juvenile fish. *J. Comp. Neurol.* **495,** 213–235 (2006).
18. Cheng, C. L. & Novales Flamarique, I. Opsin expression: new mechanism for modulating colour vision. *Nature* **428,** 279 (2004).
19. Mazzoni, E. O., Desplan, C. & Celik, A. 'One receptor' rules in sensory neurons. *Dev. Neurosci.* **26,** 388–395 (2004).
20. Applebury, M. L. *et al.* The murine cone photoreceptor: a single cone type expresses both S and M opsins with retinal spatial patterning. *Neuron* **27,** 513–523 (2000).
21. Makino, C. L. & Dodd, R. L. Multiple visual pigments in a photoreceptor of the salamander retina. *J. Gen. Physiol.* **108,** 27–34 (1996).
22. Xiao, M. & Hendrickson, A. Spatial and temporal expression of short, long/medium, or both opsins in human fetal cones. *J. Comp. Neurol.* **425,** 545–559 (2000).
23. Chandrasekaran, V. & Beckendorf, S. K. Senseless is necessary for the survival of embryonic salivary glands in *Drosophila*. *Development* **130,** 4719–4728 (2003).
24. Hock, H. *et al.* Intrinsic requirement for zinc finger transcription factor Gfi-1 in neutrophil differentiation. *Immunity* **18,** 109–120 (2003).
25. Wallis, D. *et al.* The zinc finger transcription factor Gfi1, implicated in lymphomagenesis, is required for inner ear hair cell differentiation and survival. *Development* **130,** 221–232 (2003).
26. Zheng, X. *et al.* TGF-β signaling activates steroid hormone receptor expression during neuronal remodeling in the *Drosophila* brain. *Cell* **112,** 303–315 (2003).
27. Kuo, C. T., Jan, L. Y. & Jan, Y. N. Dendrite-specific remodeling of *Drosophila* sensory neurons requires matrix metalloproteases, ubiquitin-proteasome, and ecdysone signaling. *Proc. Natl Acad. Sci. USA* **102,** 15230–15235 (2005).
28. Loi, P. K. & Tublitz, N. J. Hormonal control of transmitter plasticity in insect peptidergic neurons. I. Steroid regulation of the decline in cardioacceleratory peptide 2 (CAP2) expression. *J. Exp. Biol.* **181,** 175–194 (1993).
29. Nolo, R., Abbott, L. A. & Bellen, H. J. Senseless, a Zn finger transcription factor, is necessary and sufficient for sensory organ development in *Drosophila*. *Cell* **102,** 349–362 (2000).
30. Mikeladze-Dvali, T. *et al.* The growth regulators warts/lats and melted interact in a bistable loop to specify opposite fates in *Drosophila* R8 photoreceptors. *Cell* **122,** 775–787 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank H. Bellen, A. H. Brand, S. Britt, T. Cook, the Developmental Studies Hybridoma Bank, F. Hirth, the Kyoto Stock Center, K. Matthews, M. Mlodzik, B. Mollerau, F. Pichaud, H. Reichert, C. Thummel and J. Urban for fly stocks and antibodies. We also thank J. Blau, R. J. Johnston, A. Keene and D. Vasiliauskas for discussion and comments on the manuscript. This work was funded by grant EY013010 from the National Eye Institute/National Institutes of Health to C.D., the Swiss National Science Foundation, the Novartis Foundation and the Janggen-Pöhn Stiftung (to S.G.S.) and conducted in a facility constructed with the support of a Research Facilities Improvement Grant C06 RR-15518-01 from the National Center for Research Resources, National Institutes of Health.

Author Contributions S.G.S. performed the experimental work and analysed the data. C.D. and S.G.S. designed the experiments and wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to C.D. (cd38@nyu.edu).

LETTERS

Oligomerization of STIM1 couples ER calcium depletion to CRAC channel activation

Riina M. Luik[1]*, Bin Wang[1]*†, Murali Prakriya[1]*†, Minnie M. Wu[1] & Richard S. Lewis[1]

Ca^{2+}-release-activated Ca^{2+} (CRAC) channels generate sustained Ca^{2+} signals that are essential for a range of cell functions, including antigen-stimulated T lymphocyte activation and proliferation[1,2]. Recent studies[3] have revealed that the depletion of Ca^{2+} from the endoplasmic reticulum (ER) triggers the oligomerization of stromal interaction molecule 1 (STIM1), the ER Ca^{2+} sensor, and its redistribution to ER–plasma membrane (ER–PM) junctions[4–8] where the CRAC channel subunit ORAI1 accumulates in the plasma membrane and CRAC channels open[9–12]. However, how the loss of ER Ca^{2+} sets into motion these coordinated molecular rearrangements remains unclear. Here we define the relationships among $[Ca^{2+}]_{ER}$, STIM1 redistribution and CRAC channel activation and identify STIM1 oligomerization as the critical $[Ca^{2+}]_{ER}$-dependent event that drives store-operated Ca^{2+} entry. In human Jurkat leukaemic T cells expressing an ER-targeted Ca^{2+} indicator, CRAC channel activation and STIM1 redistribution follow the same function of $[Ca^{2+}]_{ER}$, reaching half-maximum at ~200 μM with a Hill coefficient of ~4. Because STIM1 binds only a single Ca^{2+} ion[5], the high apparent cooperativity suggests that STIM1 must first oligomerize to enable its accumulation at ER–PM junctions. To assess directly the causal role of STIM1 oligomerization in store-operated Ca^{2+} entry, we replaced the luminal Ca^{2+}-sensing domain of STIM1 with the 12-kDa FK506- and rapamycin-binding protein (FKBP12, also known as FKBP1A) or the FKBP-rapamycin binding (FRB) domain of the mammalian target of rapamycin (mTOR, also known as FRAP1). A rapamycin analogue oligomerizes the fusion proteins and causes them to accumulate at ER–PM junctions and activate CRAC channels without depleting Ca^{2+} from the ER. Thus, STIM1 oligomerization is the critical transduction event through which Ca^{2+} store depletion controls store-operated Ca^{2+} entry, acting as a switch that triggers the self-organization and activation of STIM1–ORAI1 clusters at ER–PM junctions.

The defining feature of store-operated channels is their activation in response to ER Ca^{2+} ($[Ca^{2+}]_{ER}$) depletion. However, their sensitivity to $[Ca^{2+}]_{ER}$ and the factors that determine this sensitivity have never been established, largely because of the technical difficulty of quantifying $[Ca^{2+}]_{ER}$. To address this issue, we generated a Jurkat T cell line stably expressing the Ca^{2+}-sensitive cameleon protein, YC4.2er (see Methods). YC4.2er is selectively retained in the ER, as shown by its colocalization with the resident ER protein calnexin but not with mitochondrial or Golgi markers and by its functional response to agents that deplete ER Ca^{2+} (Fig. 1a and Supplementary Fig. 1). *In situ* calibration of the YC4.2er fluorescence resonance energy transfer signal indicates a responsivity to $[Ca^{2+}]_{ER}$ in the range of ~1 μM to >1 mM (Supplementary Fig. 2).

To determine the dependence of CRAC channel activation on $[Ca^{2+}]_{ER}$, we measured CRAC current (I_{CRAC}) in perforated-patch

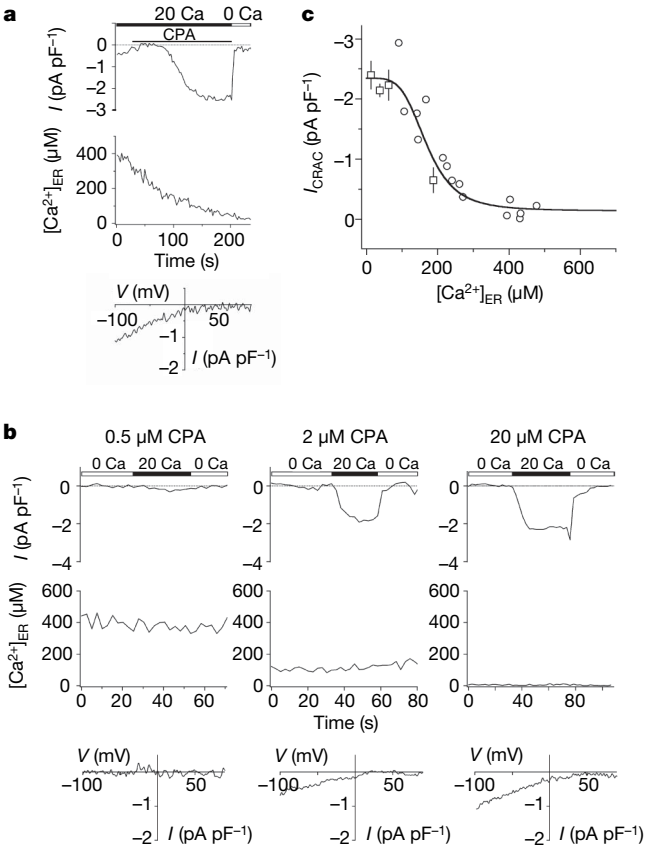


Figure 1 | The dependence of CRAC channel activation on $[Ca^{2+}]_{ER}$. Simultaneous measurements of $[Ca^{2+}]_{ER}$ and I_{CRAC} in individual Jurkat T cells. **a**, Treatment with 20 μM CPA induces an increase in I_{CRAC} (top) that follows a decrease in $[Ca^{2+}]_{ER}$ (middle) monitored with YC4.2er. The current–voltage (*I–V*) relationship shows the inward rectification typical of I_{CRAC} (bottom). In this cell, a small inward current through outwardly rectifying Cl^- channels is also present initially but disappears before I_{CRAC} is induced. Extracellular $[Ca^{2+}]$ in mM is indicated above the bars in **a** and **b**. **b**, Recordings of I_{CRAC} (top) and $[Ca^{2+}]_{ER}$ (middle) under steady-state conditions. Each cell was treated with the indicated CPA concentration for 8–15 min before recording, and CPA was maintained throughout the experiment. *I–V* relationships are typical for I_{CRAC} (bottom). **c**, Steady-state I_{CRAC} and $[Ca^{2+}]_{ER}$ are plotted for 40 cells after treatment with 0.5–20 μM CPA. A fit of the Hill equation with a $K_{1/2}$ of 169 μM and a Hill coefficient of 4.2 is superimposed on the data. Squares, mean ± s.e.m. of 3–12 cells. Circles, single cells (see Supplementary Information).

[1]Department of Molecular and Cellular Physiology, Stanford University School of Medicine, Stanford, California 94305, USA. †Present addresses: Department of Physiology, University of Texas Health Science Center at San Antonio, San Antonio, Texas 78229, USA (B.W.); Department of Molecular Pharmacology and Biological Chemistry, Feinberg School of Medicine, Northwestern University, Chicago, Illinois 60611, USA (M.P.).
*These authors contributed equally to this work.

recordings from Jurkat YC4.2er cells treated with cyclopiazonic acid (CPA), a reversible SERCA (sarco/endoplasmic reticulum Ca^{2+}-ATPase) inhibitor. CPA evokes a time-dependent decline in $[Ca^{2+}]_{ER}$ in parallel with the activation of I_{CRAC} measured in the same cell (Fig. 1a). However, because I_{CRAC} responds slowly to rapid changes of $[Ca^{2+}]_{ER}$, non-stationary measurements like these will distort estimates of the true $[Ca^{2+}]_{ER}$ dependence of the CRAC channel. For this reason, we determined instead the $[Ca^{2+}]_{ER}$–I_{CRAC} relationship under steady-state conditions by pretreating cells with 0.5–20 μM CPA for 8–15 min in the absence of extracellular Ca^{2+} to generate a range of constant $[Ca^{2+}]_{ER}$ values. This passive depletion approach also minimizes spatial variations of $[Ca^{2+}]_{ER}$, allowing the $[Ca^{2+}]_{ER}$ dependence of store-operated Ca^{2+} entry (SOCE) to be determined from whole-cell YC4.2er measurements. After re-addition of 20 mM Ca^{2+} to the bath, current was monitored during brief hyperpolarizations from the resting potential of +30–50 mV at constant $[Ca^{2+}]_{ER}$ (Fig. 1b). The current was identified as I_{CRAC} on the basis of its inwardly rectifying current–voltage relationship , extremely low current noise, and a delayed response to extracellular Ca^{2+} resulting from Ca^{2+}-dependent potentiation (Fig. 1b)[1,13]. Measurements from 40 cells show that I_{CRAC} is a steep function of $[Ca^{2+}]_{ER}$ with half-maximal activation ($K_{1/2}$) at 169 μM and a Hill coefficient of 4.2 (Fig. 1c). Interestingly, a decline of >100 μM from the resting $[Ca^{2+}]_{ER}$ of ~400 μM is required to initiate CRAC channel opening in these cells, which may help to explain how small amounts of ER Ca^{2+} can be released without activating I_{CRAC} in some cells[1].

We next addressed the source of the CRAC channel's steep dependence on $[Ca^{2+}]_{ER}$. Because STIM1 is known to be the Ca^{2+} sensor for SOCE[6,7] and its redistribution to ER–PM junctions is linked to I_{CRAC} activation[6–8,11], we measured the dependence of STIM1 redistribution on $[Ca^{2+}]_{ER}$. Exposure to 0.5–3 μM CPA for >8 min causes a partial redistribution of Cherry–STIM1 to the cell periphery, which can be seen by wide-field imaging at the cell equator (Fig. 2a). We quantified the redistribution of Cherry–STIM1 as the ratio of the mean peripheral fluorescence to the mean total fluorescence (Fig. 2b); this method gives results that agree quantitatively with total internal reflection fluorescence (TIRF) measurements of STIM1 puncta (Supplementary Fig. 3) while facilitating the separation of the cameleon and Cherry fluorescence signals (see Supplementary Information). Measurements from 41 cells show that STIM1 redistribution is a steep function of $[Ca^{2+}]_{ER}$ that closely resembles that of I_{CRAC} activation, with a $K_{1/2}$ of 187 μM and a Hill coefficient of 3.8 (Fig. 2c). The value of $K_{1/2}$ is close to the binding affinity of the recombinant EF-hand plus the sterile alpha motif (SAM) domain of STIM1 measured *in vitro* (K_d = 200–600 μM; ref. 5), consistent with its role as an ER Ca^{2+} sensor. Importantly, the close correspondence between the STIM1 and I_{CRAC} curves indicates that CRAC channels open in direct proportion to the concentration of STIM1 at ER–PM junctions and that the CRAC channel derives its highly nonlinear dependence on $[Ca^{2+}]_{ER}$ from the ER Ca^{2+} dependence of STIM1 redistribution. A recent study of HeLa cells found a similar dependence of STIM1 redistribution on $[Ca^{2+}]_{ER}$ (ref. 14). In that study, the homologue STIM2 redistributed to ER–PM junctions at higher $[Ca^{2+}]_{ER}$ ($K_{1/2}$ = 406 μM) than did STIM1 ($K_{1/2}$ = 210 μM), and it was proposed that STIM2 functions as a homeostatic ER Ca^{2+} sensor by activating ORAI1. Our findings that I_{CRAC} and STIM1 redistribution follow the same function of $[Ca^{2+}]_{ER}$ implies that in Jurkat cells STIM2 activates at most a minor fraction of endogenous CRAC channels, consistent with its low level of expression in T cells[15].

The shape of the STIM1 redistribution curve has important implications for the mechanism underlying SOCE. The Hill coefficient of ~4 shows that puncta formation is a nonlinear process with respect to $[Ca^{2+}]_{ER}$, without necessarily indicating a cooperative mechanism or that the active form of STIM1 is a tetramer. However, the high Hill coefficient implies that STIM1 puncta at ER–PM junctions do not form by the independent accretion of STIM1 monomers, which contain only a single luminal Ca^{2+}-binding site[5], but suggests instead that only oligomers of STIM1 can accumulate at these sites. There are two ways in which STIM1 is known to oligomerize. In resting cells, STIM1 self-associates with an undetermined stoichiometry by means of its cytosolic coiled-coil domains[16,17]; in addition, removal of Ca^{2+} from the EF-hand of STIM1 drives further oligomerization *in vitro*[5] and *in vivo*[4]. Store-dependent oligomerization of STIM1 occurs within seconds, slightly in advance of puncta formation, and a causal role in SOCE has been hypothesized but never tested[4,5].

To address the possible role of STIM1 oligomerization in SOCE, we adopted an approach based on rapamycin-induced protein heterodimerization[18,19]. We replaced the luminal region of Cherry–STIM1 (containing the EF-hand and SAM domains) with a tandem dimer of FK506-binding protein (FKBP12) or a variant of the FKBP-rapamycin binding domain of mTOR (FRB) to generate STIM1 chimaeras that will heterodimerize when bound to a rapamycin analogue (AP21967, or rapalogue). Given that STIM1 is known to self-associate at rest[16,17], rapalogue would thus be expected to link multimers containing FRB with those containing FKBP to form extended oligomers of STIM1 (Fig. 3a). We assayed oligomer formation in HEK293 cells expressing Cherry–FRB–STIM1 and Cherry–FKBP–STIM1 (abbreviated hereafter as F–STIM1) using blue native polyacrylamide gel electrophoresis (BN–PAGE)[20]. The >2-fold increase in apparent mass after rapalogue treatment confirms its ability to oligomerize F–STIM1, and because crosslinking of monomers would be expected to at most double the mass, indicates that the resting state of FRB–STIM1 and FKBP–STIM1 is at least a dimer (Fig. 3b).

We first examined the effects of rapalogue on the localization of F–STIM1 in Jurkat cells. Rapalogue evoked a redistribution of F–STIM1 to the cell periphery that was complete within several minutes (Fig. 3c). Quantitative analysis shows that rapalogue triggers the

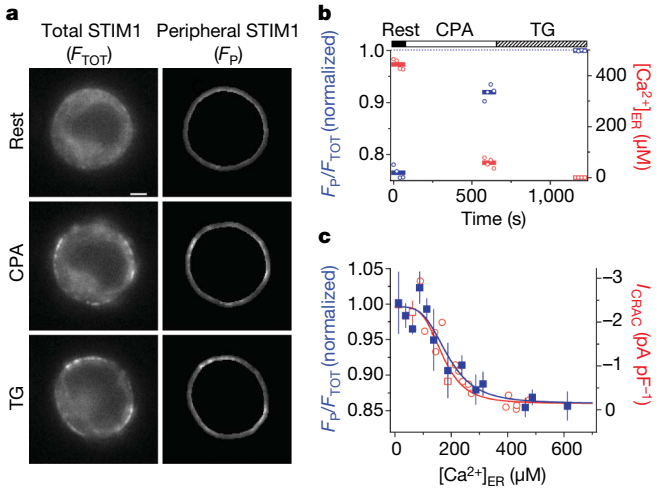


Figure 2 | The $[Ca^{2+}]_{ER}$ dependence of STIM1 redistribution determines the $[Ca^{2+}]_{ER}$-response relation of the CRAC channel. a, Wide-field epifluorescence images of a cell expressing Cherry–STIM1 at rest (top) and after store depletion with 3 μM CPA (middle) and thapsigargin (TG; bottom). The redistribution of Cherry–STIM1 in single cells was monitored as the ratio of the mean fluorescence in the most peripheral 0.5 μm of the cell (F_P, right) to the mean fluorescence of the entire cell (F_{TOT}, left). Scale bar, 2 μm. **b**, In the same cell, STIM1 redistribution is represented by F_P/F_{TOT} normalized to the maximum ratio with TG (blue). F_P/F_{TOT} increases as $[Ca]_{ER}$ (red) declines. Individual data points (open symbols) and the mean responses (bars) are shown. **c**, STIM1 redistribution (F_P/F_{TOT}, blue) is plotted against $[Ca^{2+}]_{ER}$ after treatment with 0–3 μM CPA (means ± s.e.m. of 3–4 cells; 41 cells total). A fit of the Hill equation (blue line) indicates a $K_{1/2}$ of 187 μM and a Hill coefficient of 3.8. Steady-state I_{CRAC} data fitted with the Hill equation are re-plotted from Fig. 1 (red).

redistribution of F–STIM1 as effectively as Ca^{2+}-store depletion induces the redistribution of wild-type STIM1 (Fig. 3d). When examined by TIRF microscopy, the rapalogue-driven peripheral accumulations of F–STIM1 (Fig. 3e) resemble the puncta of wild-type STIM1 that form in response to store depletion (Fig. 3f). Similar results were obtained in HEK293 cells. Rapalogue did not affect the localization of wild-type STIM1 (Fig. 3g), nor did it deplete Ca^{2+} stores (see below). Finally, rapalogue-induced F–STIM1 puncta co-localize with store-depletion-induced green fluorescent protein (GFP)–STIM1 puncta in the same cell, confirming that rapalogue causes F–STIM1 to accumulate at the same ER–PM junctions where STIM1 and ORAI1 are known to interact. Thus, we conclude that oligomerization of STIM1 is sufficient to drive the redistribution of STIM1 to ER–PM junctions.

Heterodimerization of FRB–STIM1 and FKBP–STIM1 also activates endogenous CRAC channels. Rapalogue increased the mean resting $[Ca^{2+}]_i$ in Jurkat cells expressing F–STIM1 (Cherry-positive cells) from 170 ± 11 nM (untreated; $n = 61$) to 388 ± 45 nM (Fig. 4a; $n = 45$), but did not affect $[Ca^{2+}]_i$ in untransfected Jurkat cells. The increased basal $[Ca^{2+}]_i$ was dependent on extracellular Ca^{2+} (Fig. 4a) and was inhibited by 2-aminoethyldiphenyl borate (2-APB) and low concentrations of La^{3+} (Supplementary Fig. 4), consistent with constitutive Ca^{2+} entry through open CRAC channels[1,13]. Importantly, thapsigargin (TG) released similar amounts of ER Ca^{2+} in rapalogue-pretreated and resting cells, indicating that rapalogue stimulates Ca^{2+} entry without depleting Ca^{2+} stores (Fig. 4a). Whole-cell recordings with a high-$[Ca^{2+}]$ pipette solution designed to minimize store depletion confirmed that heterodimerization of FRB–STIM1 and FKBP–STIM1 directly activates I_{CRAC}. In untreated Jurkat cells expressing F–STIM1, I_{CRAC} was negligible on breaking in to the whole-cell recording configuration and developed slowly to a small amplitude, presumably in response to partial store depletion. In contrast, in rapamycin-pretreated cells with visible puncta, large inward currents were evident immediately on breaking in (Fig. 4b) and displayed essential features of I_{CRAC}, including a dependence on extracellular Ca^{2+}, an inwardly rectifying current–voltage relationship (Fig. 4c), low current noise, rapid Ca^{2+}-dependent inactivation, and inhibition by 2-APB and La^{3+} (Supplementary Fig. 4)[13]. The mean current amplitude (2.6 ± 0.6 pA pF^{-1}, $n = 9$) was similar to that produced by Ca^{2+} store depletion in Jurkat cells overexpressing Cherry–STIM1 (ref. 11), consistent with the comparable degrees of STIM1 and F–STIM1 redistribution in response to thapsigargin or rapalogue, respectively (Fig. 3). Together, these results indicate that F–STIM1 oligomers at ER–PM junctions are fully active and provide direct evidence that the oligomerization of STIM1, independently of changes in $[Ca^{2+}]_{ER}$, is sufficient to evoke CRAC channel activation.

We have shown that STIM1 redistribution and I_{CRAC} share a steep dependence on $[Ca^{2+}]_{ER}$, and that oligomerization of F–STIM1 is sufficient to drive puncta formation and CRAC channel activation. These results define the input–output relationship of the CRAC channel and identify STIM1 oligomerization as the primary transduction event through which this relationship is determined. The EF hand and SAM domains of STIM1 seem to serve primarily to control the extent of oligomerization, considering that removal of Ca^{2+} causes a recombinant EF-SAM peptide to oligomerize *in vitro*[5], and that the FRB and FKBP modules in F–STIM1 can effectively substitute for the EF-hand and SAM domains and activate I_{CRAC} when crosslinked by rapalogue. The fact that the latter occurs without Ca^{2+} store depletion suggests that once STIM1 oligomerizes, all subsequent steps leading to SOCE occur independently of ER Ca^{2+}. Thus, we propose that the oligomerization of STIM1 acts as a switch to trigger the self-organization of STIM1 and ORAI1 complexes at ER–PM junctions and the consequent activation of CRAC channels.

How might this oligomerization 'switch' operate? In its resting state Ca^{2+}-bound STIM1 moves freely throughout the ER membrane[4], but after store depletion STIM1 oligomers accumulate in ER subregions located 10–25 nm from the PM, close enough to allow trapping by binding to targets in the PM[8]. These targets have not yet been positively identified, but suggested candidates include ORAI1 (refs 21, 22) or an associated protein[23], as well as PM phospholipids[4,24]. Once localized at ER–PM junctions, STIM1 then promotes

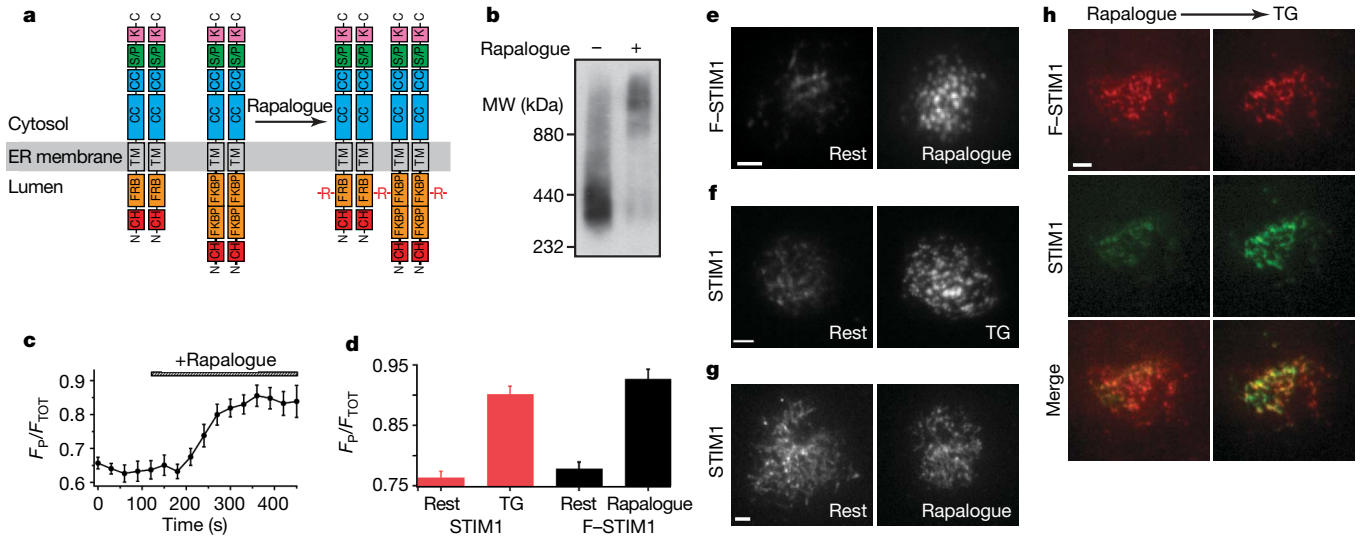


Figure 3 | STIM1 oligomerization induces the accumulation of STIM1 at ER–PM junctions. **a**, The cartoon depicts the oligomerization of F–STIM1 induced by rapalogue (R). At rest, FKBP–STIM1 and FRB–STIM1 are expected to form homo- and hetero-dimers; only intermolecular crosslinks between homodimers are shown here. Abbreviations: CC (coiled-coil), CH (monomeric Cherry), EF (EF hand), K (lysine-rich), SAM (sterile-α motif), S/P (serine–proline-rich). **b**, BN–PAGE and western blot of transiently expressed F–STIM1 harvested from HEK293 cells. Untreated (left) and rapalogue-treated (right) F–STIM1 was detected using a monoclonal anti-STIM1 antibody. MW, molecular weight. **c**, Rapalogue induces a time-dependent peripheral redistribution of F–STIM1 ($n = 10$ cells). **d**, Peripheral redistribution of Cherry–STIM1 by TG (red bars; $n = 31$ for each) and redistribution of F–STIM1 by rapalogue (black bars; $n = 39$, rest; $n = 42$, rapalogue). Values are expressed as mean ± s.e.m. in **c** and **d**. **e**–**h**, TIRF images of Jurkat cells; scale bars, 2 μm. **e**, F–STIM1 before (left) and after (right) incubation with rapalogue. **f**, Cherry–STIM1 before (left) and after (right) store depletion with TG. **g**, Cherry–STIM1 before (left) and after (right) incubation with rapalogue. **h**, F–STIM1 (top row), GFP–STIM1 (middle row) and merged images (bottom row) from a single cell after rapalogue treatment (left column) and subsequent store depletion with TG (right column). Cherry and GFP intensities are scaled to the maximal intensity of each fluorophore after TG treatment.

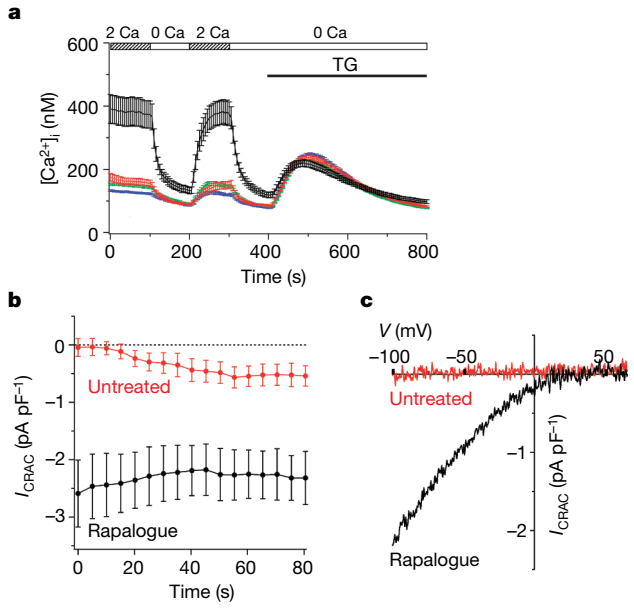


Figure 4 | STIM1 oligomerization activates Ca^{2+} entry through CRAC channels. a, In rapalogue-treated cells expressing F–STIM1 (black, $n = 45$), resting $[Ca^{2+}]_i$ is increased and sensitive to the removal of extracellular Ca^{2+}, indicating constitutive Ca^{2+} entry. In contrast, resting Ca^{2+} influx was largely absent in untreated F–STIM1-expressing cells (red, $n = 61$) and in wild-type Jurkat cells with (green, $n = 617$) or without (blue, $n = 517$) rapalogue. TG-induced Ca^{2+} release in rapalogue-treated cells was similar to that in untreated cells. **b**, I_{CRAC} development during whole-cell recording from rapalogue-treated (black, $n = 9$) and untreated (red, $n = 9$) cells expressing F–STIM1. I_{CRAC} was measured beginning within 5 s of break-in. **c**, *I–V* relations on break-in, showing the inward rectification typical of I_{CRAC} in the rapalogue-treated cell (black) and the absence of current in the untreated cell (red). Values are expressed as mean ± s.e.m. (**a**, **b**).

the accumulation of ORAI1 at apposed sites, leading to channel activation[11,12,25]. Oligomerization may promote the binding of STIM1 to its targets in two ways: an affinity-based mechanism in which a conformational change exposes a previously masked cytosolic binding domain, and an avidity-based mechanism in which clustering of the binding domains increases their local concentration at ER–PM junctions. Both of these mechanisms are likely to contribute to the assembly and function of CRAC channel complexes that constitute the final stage of the store-operated Ca^{2+} entry process.

METHODS SUMMARY

$[Ca^{2+}]_{ER}$ measurements. $[Ca]_{ER}$ was measured in a Jurkat E6-1 cell line stably expressing a modified YC4er (V68L and Q69M; refs 26–28). Cells were pre-treated with CPA (0.5–20 μM) in Ca^{2+}-free Ringer's solution for 8–15 min, and emission intensities at 485 nm and 535 nm were averaged across the cell to yield a raw emission ratio. Ratios were calibrated *in situ* for every cell as described (Supplementary Information).

Heterodimerizer experiments. To generate F–STIM1, mutant FRB and tandem FKBP sequences were substituted for the EF-SAM domain (wild-type STIM1, amino acids 35–207) in Cherry–STIM1 using plasmids provided by Ariad Pharmaceuticals. F–STIM1 was crosslinked using 1 μM rapalogue (AP21967, Ariad Pharmaceuticals). Unless indicated otherwise, cells were pre-incubated in full medium at 37 °C for 30 min, with or without rapalogue, and subsequent measurements were performed at 22–25 °C in standard Ringer's solutions. Time-lapse imaging was performed at 37 °C in full medium with or without rapalogue. Only cells with ~3–10% of the fluorescence of the brightest cells in each experiment were analysed. BN–PAGE was performed essentially as described[20]. A monoclonal antibody against the STIM1 carboxy terminus (1:250; Abnova) and an alkaline-phosphatase-conjugated secondary antibody (1:30,000; Sigma) were used for western blotting.

Perforated-patch and whole-cell recording. I_{CRAC} in YC4.2er cells (Fig. 1) was recorded in the perforated-patch configuration[29] with 20 mM extracellular Ca^{2+}, using a stimulus of a 50-ms step to −100 mV followed by a ramp from −100 to +100 mV, delivered from a holding potential of +30 or +50 mV. Whole-cell recording of I_{CRAC} (ref. 30; Fig. 4) was performed with 20 mM Ca^{2+} Ringer's solution, with stimuli consisting of a 100-ms step to −112 mV followed by a 100-ms voltage ramp from −112 to +88 mV applied from the holding potential of +38 mV beginning within 5 s of break-in.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 21 January; accepted 8 May 2008.
Published online 2 July 2008.

1. Parekh, A. B. & Putney, J. W. Jr. Store-operated calcium channels. *Physiol. Rev.* **85,** 757–810 (2005).
2. Feske, S. Calcium signalling in lymphocyte activation and disease. *Nature Rev. Immunol.* **7,** 690–702 (2007).
3. Wu, M. M., Luik, R. M. & Lewis, R. S. Some assembly required: constructing the elementary units of store-operated Ca^{2+} entry. *Cell Calcium* **42,** 163–172 (2007).
4. Liou, J., Fivaz, M., Inoue, T. & Meyer, T. Live-cell imaging reveals sequential oligomerization and local plasma membrane targeting of stromal interaction molecule 1 after Ca^{2+} store depletion. *Proc. Natl Acad. Sci. USA* **104,** 9301–9306 (2007).
5. Stathopulos, P. B., Li, G. Y., Plevin, M. J., Ames, J. B. & Ikura, M. Stored Ca^{2+} depletion-induced oligomerization of STIM1 via the EF–SAM region: an initiation mechanism for capacitive Ca^{2+} entry. *J. Biol. Chem.* **281,** 35855–35862 (2006).
6. Zhang, S. L. *et al.* STIM1 is a Ca^{2+} sensor that activates CRAC channels and migrates from the Ca^{2+} store to the plasma membrane. *Nature* **437,** 902–905 (2005).
7. Liou, J. *et al.* STIM is a Ca^{2+} sensor essential for Ca^{2+}-store-depletion-triggered Ca^{2+} influx. *Curr. Biol.* **15,** 1235–1241 (2005).
8. Wu, M. M., Buchanan, J., Luik, R. M. & Lewis, R. S. Ca^{2+} store depletion causes STIM1 to accumulate in ER regions closely associated with the plasma membrane. *J. Cell Biol.* **174,** 803–813 (2006).
9. Prakriya, M. *et al.* Orai1 is an essential pore subunit of the CRAC channel. *Nature* **443,** 230–233 (2006).
10. Vig, M. *et al.* CRACM1 multimers form the ion-selective pore of the CRAC channel. *Curr. Biol.* **16,** 2073–2079 (2006).
11. Luik, R. M., Wu, M. M., Buchanan, J. & Lewis, R. S. The elementary unit of store-operated Ca^{2+} entry: local activation of CRAC channels by STIM1 at ER–plasma membrane junctions. *J. Cell Biol.* **174,** 815–825 (2006).
12. Xu, P. *et al.* Aggregation of STIM1 underneath the plasma membrane induces clustering of Orai1. *Biochem. Biophys. Res. Commun.* **350,** 969–976 (2006).
13. Prakriya, M. & Lewis, R. S. CRAC channels: activation, permeation, and the search for a molecular identity. *Cell Calcium* **33,** 311–321 (2003).
14. Brandman, O., Liou, J., Park, W. S. & Meyer, T. STIM2 is a feedback regulator that stabilizes basal cytosolic and endoplasmic reticulum Ca^{2+} levels. *Cell* **131,** 1327–1339 (2007).
15. Oh-Hora, M. *et al.* Dual functions for the endoplasmic reticulum calcium sensors STIM1 and STIM2 in T cell activation and tolerance. *Nature Immunol.* **9,** 432–443 (2008).
16. Baba, Y. *et al.* Coupling of STIM1 to store-operated Ca^{2+} entry through its constitutive and inducible movement in the endoplasmic reticulum. *Proc. Natl Acad. Sci. USA* **103,** 16704–16709 (2006).
17. Williams, R. T. *et al.* Stromal interaction molecule 1 (STIM1), a transmembrane protein with growth suppressor activity, contains an extracellular SAM domain modified by *N*-linked glycosylation. *Biochim. Biophys. Acta* **1596,** 131–137 (2002).
18. Bayle, J. H. *et al.* Rapamycin analogs with differential binding specificity permit orthogonal control of protein activity. *Chem. Biol.* **13,** 99–107 (2006).
19. Varnai, P., Thyagarajan, B., Rohacs, T. & Balla, T. Rapidly inducible changes in phosphatidylinositol 4,5-bisphosphate levels influence multiple regulatory functions of the lipid in intact living cells. *J. Cell Biol.* **175,** 377–382 (2006).
20. Schägger, H. & von Jagow, G. Blue native electrophoresis for isolation of membrane protein complexes in enzymatically active form. *Anal. Biochem.* **199,** 223–231 (1991).
21. Muik, M. *et al.* Dynamic coupling of the putative coiled-coil domain of ORAI1 with STIM1 mediates ORAI1 channel activation. *J. Biol. Chem.* **283,** 8014–8022 (2008).
22. Yeromin, A. V. *et al.* Molecular identification of the CRAC channel by altered ion selectivity in a mutant of Orai. *Nature* **443,** 226–229 (2006).
23. Varnai, P., Toth, B., Toth, D. J., Hunyady, L. & Balla, T. Visualization and manipulation of plasma membrane–endoplasmic reticulum contact sites indicates the presence of additional molecular components within the STIM1–Orai1 complex. *J. Biol. Chem.* **282,** 29678–29690 (2007).
24. Huang, G. N. *et al.* STIM1 carboxyl-terminus activates native SOC, I_{crac} and TRPC1 channels. *Nature Cell Biol.* **8,** 1003–1010 (2006).
25. Li, Z. *et al.* Mapping the interacting domains of STIM1 and Orai1 in Ca^{2+} release-activated Ca^{2+} channel activation. *J. Biol. Chem.* **282,** 29448–29456 (2007).
26. Miyawaki, A., Griesbeck, O., Heim, R. & Tsien, R. Y. Dynamic and quantitative Ca^{2+} measurements using improved cameleons. *Proc. Natl Acad. Sci. USA* **96,** 2135–2140 (1999).

27. Miyawaki, A. *et al.* Fluorescent indicators for Ca^{2+} based on green fluorescent proteins and calmodulin. *Nature* **388,** 882–887 (1997).
28. Griesbeck, O., Baird, G. S., Campbell, R. E., Zacharias, D. A. & Tsien, R. Y. Reducing the environmental sensitivity of yellow fluorescent protein. Mechanism and applications. *J. Biol. Chem.* **276,** 29188–29194 (2001).
29. Bautista, D. M., Hoth, M. & Lewis, R. S. Enhancement of calcium signalling dynamics and stability by delayed modulation of the plasma-membrane calcium–ATPase in human T cells. *J. Physiol. (Lond.)* **541,** 877–894 (2002).
30. Zweifach, A. & Lewis, R. S. Slow calcium-dependent inactivation of depletion-activated calcium current. Store-dependent and -independent mechanisms. *J. Biol. Chem.* **270,** 14445–14451 (1995).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank N. Bhakta and D. Bautista for assistance and advice during the initial phase of these studies, R. Tsien for the gift of cameleon YC4er, P. Bacchawat for advice on BN–PAGE, and R. Dolmetsch for comments on the manuscript. This work was supported by a grant from the National Institutes of Health (NIH) and the Mathers Charitable Foundation.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to R.S.L. (rslewis@stanford.edu).

METHODS

Cells, solutions and reagents. Jurkat E6-1 human leukaemic T cells (American Type Culture Collection) and Jurkat YC4.2er cell lines were maintained as described previously[31]. Unless indicated otherwise, experiments were performed at 22–25 °C after cells were attached to poly-D-lysine-coated coverslip chambers and were bathed in Ringer's solution containing (in mM): 155 NaCl, 4.5 KCl, 2 or 20 $CaCl_2$, 1 $MgCl_2$, 10 D-glucose and 5 Na-HEPES (pH 7.4). For Ca^{2+}-free Ringer's solution, $CaCl_2$ was replaced with 1 mM EGTA plus 2 mM $MgCl_2$. To deplete Ca^{2+} stores, cells were exposed to Ca^{2+}-free Ringer's solution supplemented with CPA (0.5–20 μM) or thapsigargin (1 μM). All salts and chemicals were from Sigma unless otherwise stated. Thapsigargin was purchased from LC Laboratories, ionomycin and digitonin were from EMD Biosciences, Inc., monoclonal antibody OKT3 was from eBioscience and Fura-2/AM and Mitotracker Red were purchased from Invitrogen. AP21967 (rapalogue) was provided by Ariad Pharmaceuticals (www.ariad.com/regulationkits).

Plasmids and transfection. Point mutations V68L and Q69M were introduced into the original YC4er (provided by R. Tsien) to generate YC4.2er (refs 26–28). Jurkat E6–1 cells were transfected with YC4.2er by electroporation and selected for stable expression with G418. A monoclonal line with the highest cameleon expression and normal SOCE was used. Construction of Cherry–STIM1 and GFP–myc–ORAI1 were described previously[11].

For the construction of F–STIM1, mutant FRB (pC_4-R_HE) and tandem FKBP (pC_4M-F2E) plasmids were provided by Ariad Pharmaceuticals. The provided variant of FRB binds an analogue of rapamycin (AP21976, or rapalogue) that fails to bind native mTOR, enabling heterodimerization of FRB and FKBP while avoiding potential side effects from inhibition of mTOR kinase activity[18]. FRB- and tandem FKBP-containing STIM1 plasmids were made by engineering two unique restriction sites, a NheI site after Cherry and a MluI site after the SAM domain, into Cherry–STIM1 using site-directed mutagenesis (Quikchange XL, Stratagene). The plasmid was then digested with NheI and MluI to remove the EF-hand and SAM domains (amino acids 35–207 in the wild-type STIM1 sequence). FRB was amplified from pC_4-R_HE using the primers (5′–3′) TGATTAGCTAGCGGTGCTGGTGCTGGTGCTGGTGCTGGTGCTGGTATCCTCTGGCATGAG and (5′–3′) CGACGAATCTCAAAGGGAGCAGGAGCAGGAGCAGGAGCAGGAGCAGGAACGCGTGTAATT to append 11 amino acid linkers (GAGAGAGAGAG) flanked by NheI or MluI restriction sites, respectively. Tandem FKBP (2×FKBP) was amplified from pC_4M-F2E using the primers (5′–3′) TGATTAGCTAGCGGTGCTGGTGCTGGTGCTGGTGCTGGTGCTGGTGCTGGTGGAGTGCAGGTGGAA and (5′–3′) CTGCTGAAGCTGGAGGGAGCAGGAGCAGGAGCAGGAGCAGGAGCAGGAACGCGTGTAATT. Amplified, the FRB and 2×FKBP were separately ligated into NheI- and MluI-digested Cherry–STIM1. The introduction of the NheI site introduced a premature STOP codon, which was removed by site-directed mutagenesis.

FRB–STIM1 and FKBP–STIM1 were transiently transfected into Jurkat E6–1 or Jurkat YC4.2er cells by electroporation as described[8], or into HEK293 cells by lipofection following the manufacturer's protocol (Invitrogen) using 0.5 μg of each construct. Cells were studied 48–72 h after transfection.

Fluorescence microscopy. Wide-field epifluorescence and TIRF microscopy was performed essentially as described using a Zeiss Axiovert 200M microscope[8]. For wide-field epifluorescence microscopy of YC4.2er, cells were excited at 420 nm and dual emission ratios were collected using a 455 DCLP filter cube (Chroma) and by rapidly alternating D485/40 and D535/30 emission filters with a filter changer (Lambda 10-2, Sutter Instruments) positioned at the exit port of the microscope. All images were acquired with a cooled CCD camera (ORCA-ER, Hamamatsu) using 2 × 2 binning (GFP, Cherry, YC4.2er) or 4 × 4 binning (YC4.2er when co-expressed with Cherry).

Fluorescence resonance energy transfer measurements of $[Ca^{2+}]_{ER}$. Background-corrected emission intensities at 535 nm and 485 nm were averaged across the cell to yield a raw F_{535}/F_{485} emission ratio. At the end of every experiment, *in situ* calibration of YC4.2er cells was performed by adding digitonin (50–75 μg ml^{-1}) and ionomycin (10 μM) to permeabilize the plasma membrane while leaving intracellular organelles intact and to equilibrate Ca^{2+} across the ER membrane (Supplementary Fig. 2). In the presence of ionomycin and digitonin, two standard F_{535}/F_{485} ratios (R_1 and R_2) were obtained after exposing the cells to (in mM): 75 K aspartate, 60 KCl, 1 $MgCl_2$, and either 10 EGTA (for R_1) or 20 $CaCl_2$ (for R_2). For every cell, raw F_{535}/F_{485} ratios (R) were normalized using these standard solutions:

$$R_{norm} = (R - R_1)/(R_2 - R_1)$$

where R_{norm} is the normalized YC4.2er ratio[27]. R_{norm} was converted to $[Ca^{2+}]_{ER}$ using a calibration curve determined separately.

A complete calibration curve for YC4.2er (Supplementary Fig. 2) was generated by exposing cells to solutions with various Ca^{2+} concentrations (listed in Supplementary Table 1) after permeabilization with digitonin and ionomycin as described above. The concentrations of K aspartate and KCl were adjusted to maintain constant $[Cl^-]$ and osmolarity. Free $[Ca^{2+}]$ was calculated with MaxChelator software (http://maxchelator.stanford.edu) and subsequently adjusted using a calibrated Ca^{2+}-sensitive electrode (Orion Research Inc.). Normalized F_{535}/F_{485} ratios were calculated as described above and plotted against free $[Ca^{2+}]$, and this relationship was fitted with the following equation using IgorPro (Wavemetrics):

$$[Ca^{2+}]_{ER} = K_d[(R_{norm} - R_{min})/(R_{max} - R_{norm})]^{1/n}$$

where K_d (819 μM) is the apparent dissociation constant, n is 0.54, R_{min} is 0.193 and R_{max} is 1.134. YC4.2er is reported to have an additional, high-affinity binding site for $[Ca^{2+}]_{ER}$ (K_d = 83 nM; ref. 27); the calibration curve was fitted only to the lower affinity site, which reported $[Ca^{2+}]_{ER}$ from ~400 nM to >1 mM (Supplementary Fig. 2b). Thus, in the few cells where R_{norm} fell below the R_{min} for this low-affinity binding site, $[Ca^{2+}]_{ER}$ was assigned a value of 400 nM.

Immunocytochemistry. Jurkat YC4.2er cells were attached to poly-D-lysine-coated glass coverslips and fixed in 4% fresh paraformaldehyde at 22–25 °C for 15 min. For staining with polyclonal anti-calnexin antibody (Stressgen Biotechnologies Corp.), cells were permeabilized for 5 min in cold methanol at –20 °C. For staining with monoclonal anti-golgin-97 (Molecular Probes), cells were permeabilized in 0.5% Triton X-100 in phosphate-buffered saline containing 0.2% bovine serum albumin (PBS/BSA) for 5 min at 22–25 °C. After permeabilization, cells were rinsed three times with PBS and incubated in blocking buffer containing 20 mM glycine, 75 mM NH_4Cl, 0.2% BSA and 1% goat serum in PBS for 1 h. Cells were then incubated with primary antibodies (1:1,000 dilution in blocking buffer) for an additional 1 h and rinsed three times with blocking buffer to remove unbound antibody. Alexa 594-conjugated secondary antibody (Molecular Probes; 1:1,000 dilution in PBS/BSA) was applied for 45 min, followed by three rinses with PBS/BSA. Coverslips were mounted in Vectashield (Vector Laboratories) and viewed with a Zeiss Axiovert 200M microscope (×40 NA1.3 oil), or a Molecular Dynamics Multiprobes 2010 confocal microscope (×40 NA1.3 oil).

Quantification of STIM1 redistribution. A binary mask of the cell periphery was applied to each background-corrected Cherry image by drawing a polygonal region with a width of 0.5 μm around the edge of the cell. The extent of Cherry redistribution was assessed as the mean intensity within the masked region (F_P) divided by the intensity averaged across the whole cell (F_{TOT}). In TIRF experiments, background-corrected Cherry images were thresholded to exclude remaining background fluorescence, and intensity was measured within an outline drawn around the cell footprint in the first frame of each time-lapse series. To compensate for any cell-to-cell differences in the abundance of peripheral ER, changes in redistribution measured by TIRF or wide-field epifluorescence microscopy were normalized to the maximal response obtained in each cell after store depletion with thapsigargin. Mean values of normalized F_P/F_{TOT} or TIRF intensities were calculated from 3–4 images for each data point.

Heterodimerizer experiments. All experiments were performed using 1 μM AP21967. For time-lapse imaging of F–STIM1 redistribution in response to AP21967, cells were imaged using wide-field epifluorescence at 37 °C in full medium. TIRF measurements of F–STIM1 were made at 22–25 °C in 2 mM Ca^{2+} Ringer's solution, before and after treatment with AP21967 for 30 min in full medium at 37 °C on the microscope stage; untreated cells were incubated in full medium alone under the same conditions. Otherwise, cells were pretreated for 30 min in full medium with (treated cells) or without (untreated cells) AP21967 at 37 °C in a CO_2 incubator before imaging at 22–25 °C. Cells overexpressing STIM1 or F–STIM1 at high levels often displayed puncta in the resting state, and correspondingly increased resting $[Ca^{2+}]_i$. This effect of overexpression has also been reported by others for STIM1 (ref. 7). To avoid these effects of overexpression, we restricted analysis to cells with ~3–10% of the fluorescence of the brightest cells in each experiment.

BN–PAGE and western blot analysis. BN–PAGE was performed using the NativePAGE Novex Bis-Tris gel system (Invitrogen) according to the manufacturer's instructions. In brief, 10^7 HEK293 cells expressing F–STIM1 were solubilized using NativePAGE sample buffer supplemented with 1% *n*-dodecyl-β-D-maltoside. For the rapalogue condition, cells were incubated with 1 μM AP21967 for 30 min at 37 °C before solubilization, and 1 μM AP21967 was included in the sample buffer. Coomassie G-250 was added to samples, and 0.5–1% of the sample was loaded onto the NativePAGE Novex 4–16% Bis-Tris gel. Proteins bound to Coomassie G-250 were transferred electrophoretically to Hybond-P membrane (GE Healthcare) in Tris/Glycine buffer (BioRad) for 14 h at 60 mA,

and then fixed with 10% acetic acid and de-stained with methanol. Membranes were blocked with 5% skimmed milk powder in Tris/NaCl buffer (0.05 M Tris, pH 7.4, 0.15 M NaCl, 0.05% Tween 20) and incubated for 8 h at 4 °C with a monoclonal antibody against the STIM1 C terminus (1:250, Abnova). The membrane was washed with Tris/NaCl buffer and then incubated with alkaline phosphatase-conjugated secondary antibody (1:30,000; Sigma) for 1 h at 22–25 °C. After subsequent washing, alkaline phosphatase was detected using Lumi-Phos substrate (Pierce).

Cytosolic $[Ca^{2+}]_i$ measurements. Video microscopic measurements of $[Ca^{2+}]_i$ were performed as described previously[29].

Electrophysiology. Patch-clamp recording was conducted in the standard whole-cell and perforated-patch configurations as previously described[29,30]. The internal solution for whole-cell recording contained (in mM): 140 Cs aspartate, 5 $MgCl_2$, 0.5 $CaCl_2$, 1.2 EGTA and 10 HEPES (pH 7.2 with CsOH). The free $[Ca^{2+}]$ of this solution was calculated to be 146 nM using MaxChelator. For perforated-patch experiments, the internal solution contained (in mM): 115 Cs aspartate, 1 $CaCl_2$, 5 $MgCl_2$, 10 NaCl, 10 HEPES and 100 μg ml^{-1} amphotericin B (pH 7.2 with CsOH). All data were leak-subtracted using currents collected in Ca^{2+}-free Ringer's solution. For whole-cell recording of I_{CRAC} in cells expressing F–STIM1, resting cells were perfused with 20 mM Ca^{2+} Ringer's solution in the cell-attached configuration, and voltage stimuli consisting of a 100-ms step to –112 mV followed by a 100-ms voltage ramp from –112 to +88 mV were applied from the holding potential of +38 mV every 2–5 s, beginning within 5 s of break-in. I_{CRAC} was measured as a 10-ms average at the end of 100-ms pulses to −112 mV, and spontaneous I_{CRAC} was measured from the first voltage step to −112 mV after break-in.

Measuring I_{CRAC} and $[Ca^{2+}]_{ER}$. After 8–15 min pretreatment with CPA (0.5–20 μM) in Ca^{2+}-free Ringer's solution, the perforated-patch configuration was established with YC4.2er Jurkat cells. I_{CRAC} was measured after perfusing the cells with a 20 mM Ca^{2+} Ringer's solution, using a step/ramp stimulus consisting of a 50-ms step to −100 mV followed by a voltage ramp from –100 to +100 mV delivered every 2.5 s from the holding potential of +30 or +50 mV. $[Ca^{2+}]_{ER}$ was measured simultaneously with I_{CRAC} by exciting at 440 ± 10 nm (Chroma) for 40 ms every 2.5 s through a Nikon Fluor ×40 objective (NA 1.3). The emissions from YC4.2er at 485 ± 12.5 nm and 535 ± 12.5 nm were collected simultaneously from an area slightly larger than the cell using two photomultipliers (HC120 05-MOD; Hamamatsu). YC4.2er ratios were calibrated *in situ* as described above.

Image and data analysis. Image analysis was performed using ImageJ software (National Institutes of Health). The dependence of STIM1 redistribution and I_{CRAC} (y) on $[Ca^{2+}]_{ER}$ was described by the Hill equation:

$$y = y_{min} + (y_{max} - y_{min})/(1 + (K_{1/2}/[Ca^{2+}]_{ER})^{n_H})$$

where y_{max} and y_{min} are the maximal and minimal values of STIM1 redistribution or I_{CRAC}, $K_{1/2}$ is the $[Ca^{2+}]_{ER}$ at which these values are half-maximal, and the Hill coefficient, n_H, is a measure of the steepness of the relationship. A curve was fitted by nonlinear least squares in Igor Pro to the entire collection of single-cell data using this equation. For display purposes in Figs 1 and 2, the single-cell responses within 25-μM bins of $[Ca^{2+}]_{ER}$ were averaged and plotted as mean values ± s.e.m. Because of the difficulty of obtaining large numbers of cells in the perforated-patch experiments, some $[Ca^{2+}]_{ER}$ bins contained fewer than three cells; for these bins, I_{CRAC} in single cells is plotted instead (Fig. 1).

31. Prakriya, M. & Lewis, R. S. Potentiation and inhibition of Ca^{2+} release-activated Ca^{2+} channels by 2-aminoethyldiphenyl borate (2-APB) occurs independently of IP_3 receptors. *J. Physiol. (Lond.)* **536**, 3–19 (2001).

LETTERS

Saccharomyces cerevisiae ATM orthologue suppresses break-induced chromosome translocations

Kihoon Lee[1], Yu Zhang[1] & Sang Eun Lee[1]

Chromosome translocations are frequently associated with many types of blood-related cancers and childhood sarcomas. Detection of chromosome translocations assists in diagnosis, treatment and prognosis of these diseases[1]; however, despite their importance to such diseases, the molecular mechanisms leading to chromosome translocations are not well understood. The available evidence indicates a role for non-homologous end joining (NHEJ) of DNA double-strand breaks (DSBs) in their origin[1–3]. Here we develop a yeast-based system that induces a reciprocal chromosome translocation by formation and ligation of breaks on two different chromosomes. We show that interchromosomal end joining is efficiently suppressed by the Tel1- and Mre11–Rad50–Xrs2-dependent pathway; this is distinct from the role of Tel1 in telomeric integrity and from Mec1- and Tel1-dependent checkpoint controls. Suppression of DSB-induced chromosome translocations depends on the kinase activity of Tel1 and Dun1, and the damage-induced phosphorylation of Sae2 and histone H2AX proteins. Tel1- and Sae2-dependent tethering and promotion of 5′ to 3′ degradation of broken chromosome ends discourage error-prone NHEJ and interchromosomal NHEJ, preserving chromosome integrity on DNA damage. Our results indicate that, like human ATM, Tel1 serves as a key regulator for chromosome integrity in the pathway that reduces the risk for DSB-induced chromosome translocations, and are probably pertinent to the oncogenic chromosome translocations in ATM-deficient cells.

The analysis of breakpoints from many chromosome translocations indicates that DSBs and rejoining by NHEJ are involved in the formation of recurring chromosome translocations[1–4]. To define the molecular basis of NHEJ-mediated chromosome translocations, a haploid yeast strain, SLY60, was constructed that carried a galactose-inducible HO endonuclease gene at the *ADE3* locus and two HO cleavage sites located on different chromosomes, each joined to an artificial intron (AI), as well as to the 5′ or 3′ half of the *URA3* gene (Fig. 1a). The strain is a uracil auxotroph. Chromosome translocation restores the complete *URA3* gene and the intron sequence to become Ura+. The flexibility of intronic sequences will permit cells with minor junctional modifications during translocation to grow in uracil-deficient medium. Strains with a single HO cleavage at either chromosome confirmed efficient DSB induction and repair at both sites (Supplementary Fig. 1). Parallel plasmid-based end joining assays independently measured end-joining efficiency for each strain (Supplementary Fig. 2).

We induced the HO endonuclease by adding galactose to the medium for 1 h (Fig. 1b). We then plated cells onto glucose-containing media, which shuts off HO expression, both with and without uracil. Growth on the uracil-containing media was used to measure the survival frequency after two DSBs by NHEJ, whereas growth without uracil was used to calculate the chromosome translocation frequency. Additionally, we induced HO expression by plating cells directly onto galactose-containing media, supplemented with or without uracil (Fig. 1c). Expression of HO for 1 h assessed chromosome translocations by simple re-ligation, whereas persistent HO induction measured chromosome translocations by imprecise (mutagenic) end joining[5].

In both cases, two contemporaneous DSBs on different chromosomes generated the Ura+ survivors that represent chromosome translocations. Amplification of junction sequences or electrophoretic karyotyping of Ura+ survivors verified reciprocal chromosome translocations (Supplementary Figs 3 and 5). Analysis of the breakpoint sequences supports the role of NHEJ in chromosome translocations, as all involve either perfect re-ligation (with 1 h HO expression) or imprecise end-joining (with persistent HO expression) (Supplementary Fig. 4)[5]. The role of Ku-dependent NHEJ in chromosome translocations is further ascertained by the fact that deletion of *yKu70*, *DNL4*, *MRE11*, *RAD50*, or *XRS2* abolished the formation of Ura+ survivors (Fig. 1b, c, Supplementary Table 2). Overall, chromosome translocations account for less than 10% of the total end-repair events (Supplementary Table 2). Therefore, Ku-dependent NHEJ has a strong bias against the joining of ends that leads to reciprocal chromosomal translocation. The results are in contrast to that of single-strand annealing, where annealing between two different chromosomes occurs as frequently as that between the same chromosomes[6].

In mammals, mutations in the *Atm* gene lead to frequent chromosome translocations and predisposition to lymphoma[7]. In yeast, deletion of both *MEC1* (a yeast ATR homologue) and *TEL1* (a yeast ATM homologue) increased gross chromosomal rearrangements ~10,000-fold, among which most were chromosome translocations[4]. We found that the deletion of *TEL1* increased the frequency of NHEJ-mediated chromosome translocations fourfold by simple re-ligation and 11-fold by imprecise NHEJ (Fig. 1b, c). The absence of *TEL1* also increased imprecise NHEJ (3.8-fold), resulting in mostly small deletions (–ACA, ~100-fold) (Supplementary Fig. 4). The absence of *TEL1* did not significantly alter the rate of NHEJ (Supplementary Fig. 5). Deletion of *yKu70* abolished chromosome translocations in *tel1*Δ cells (Fig. 1b, 1c). These results support the role of *TEL1* in suppressing mutagenic NHEJ and chromosome translocations.

The two most well known phenotypes of *TEL1*-deleted cells are short telomeres[8] and the lack of damage-inducible cell cycle arrest at G2/M when *MEC1* is inactivated[9]. *tel1-11* cells are defective in DSB signalling functions, but show near-wild-type levels of telomere homeostasis at the non-permissive temperature (37 °C)[10]. We found that *tel1-11* elevates the frequency of NHEJ-mediated chromosome translocations as much as *tel1*Δ does (Fig. 1b, c). Furthermore, deletion of *TLC1*, a key telomerase subunit that markedly reduces telomere length and integrity[11], did not increase the frequency of chromosome translocations. Therefore, the increase in NHEJ-mediated chromosome translocations in *tel1*Δ cells is not a result of unstable or short telomeres.

[1]Department of Molecular Medicine and Institute of Biotechnology, University of Texas Health Science Center at San Antonio, San Antonio, Texas 78245, USA.

Tel1 has a minor role in damage-induced checkpoint arrest[8,9]. Mec1 has a more visible role in G1, S and G2/M checkpoints[12]. To investigate whether the lack of damage-induced cell cycle arrest might produce more chromosome translocations in *tel1*Δ cells, we tested whether a *MEC1* deletion (and *SML1* deletion to permit *mec1*Δ cells to survive)[13] mutant also increases the frequency of chromosome translocations. We found that *MEC1* deletion only slightly increased the frequency of chromosome translocations (Fig. 1b, c). Similarly, deletion of *Saccharomyces RAD9* and *RAD17* (an adaptor for Chk2 activation and a component of the 9-1-1 complex, respectively[12]) caused a modest increase in chromosomal translocations (Fig. 1b, c). We also found that the frequency of chromosome translocations in *mec1*Δ *tel1*Δ *sml1*Δ cells is largely indistinguishable from that of *tel1*Δ cells (Fig. 1b, c). Thus, suppression of chromosome translocations is distinct from damage-induced checkpoint regulation, and Mec1 contributes very little (if any) to this function.

Activation of Tel1 depends on the Mre11 complex (Mre11–Rad50–Xrs2) and unprocessed DNA breaks[9]. Even if Sae2 facilitates end processing and inhibits Tel1 function[9], DNA repair by the Mre11 complex or Sae2 is enhanced by Tel1 (refs 9, 14, 15). Therefore, we examined the effect of mutations in *SAE2* or *MRE11* on the frequency of chromosome translocations. We found that deletion of *SAE2* increased the frequency of chromosome translocations slightly higher than that in *tel1*Δ cells (Fig. 1b, c). Mutants in which both *SAE2* and *TEL1* were deleted exhibited no more chromosome translocations than each single-gene deletion mutant, suggesting that they belong to the same pathway of suppressing chromosome translocations (Fig. 1b, c). We also found that the nuclease-defective variants of Mre11 (*mre11-H125N*)[16] or Rad50 (*rad50S*)[17] increased chromosome translocations by 30- to 44-fold relative to wild type (Fig. 1c). The chromosome translocations in all of these mutants rely on NHEJ, because deletion of *yKu70* or *DNL4* abolished formation of chromosome translocations (Supplementary Table 2 and data not shown). Junction sequences and kinetics of NHEJ in these mutants further support the role of these genes in the same pathway with *TEL1* (Supplementary Figs 4 and 5 and data not shown). Collectively, these results suggest that the Tel1-, Mre11- and Sae2-dependent pathway suppresses NHEJ-mediated chromosome translocations in yeast.

Tel1 probably suppresses chromosome translocations via phosphorylation of target proteins. Indeed, expression of a Tel1 kinase-dead mutant (*tel1-kd*)[8] increased chromosome translocations just as the *TEL1* deletion does, whereas deletion of *CHK1* or expression of *rad53-D339A* kinase-dead mutant[18] did not change chromosome translocation efficiency (Fig. 1b, c). *DUN1* encodes a protein kinase that regulates damage-inducible gene expression. Rad53 phosphorylates Dun1 on DNA damage, even though not every Dun1 function overlaps with those of Rad53. Notably, Mec1 (and probably Tel1 also) can directly phosphorylate Dun1 *in vitro*[19]. Therefore, the Rad53-independent functions of Dun1 may rely on phosphorylation by Tel1. We found that deletion of *DUN1* or expression of *dun1-kd* (ref. 18) increases chromosome translocations by imprecise NHEJ (Fig. 1b, c). The frequency of chromosome translocations in *dun1*Δ *tel1*Δ cells is identical to that of *tel1*Δ cells, indicating that Dun1 suppresses NHEJ-mediated chromosome translocations as a kinase downstream to Tel1.

The low level of chromosome translocations attributable to religation in *chk1*Δ, *dun1*Δ, or *rad53-D339A* mutants suggests that Tel1 suppresses chromosome translocations by virtue of its kinase activity on other targets. The Sae2 and H2AX proteins are post-translationally modified by either Te1l or Mec1 kinases in a damage-dependent manner[20,21]. We asked whether phosphorylation of H2AX or Sae2 suppresses chromosome translocations by measuring chromosome translocation frequency in yeast strains carrying the *sae2*^{2,5,6,8,9} (*sae2* mutant with substitutions of all five (S/T)Q to A) and the *hta1-S129A hta2-S129A* mutations that block their damage-induced phosphorylation[20,21]. Expression of *sae2*^{2,5,6,8,9} or *hta1-S129A hta2-S129A* elevated the frequency of NHEJ-mediated chromosome translocations (Fig. 1b, c). Furthermore, chromosomal translocations in *sae2*^{2,5,6,8,9} or *hta1-S129A hta2-S129A* strains did not increase further on deletion of *TEL1* or *SAE2* (Fig. 1b, c). These results suggest that Tel1-dependent phosphorylation of Dun1, H2AX and Sae2 suppresses chromosome translocation of HO cleaved ends.

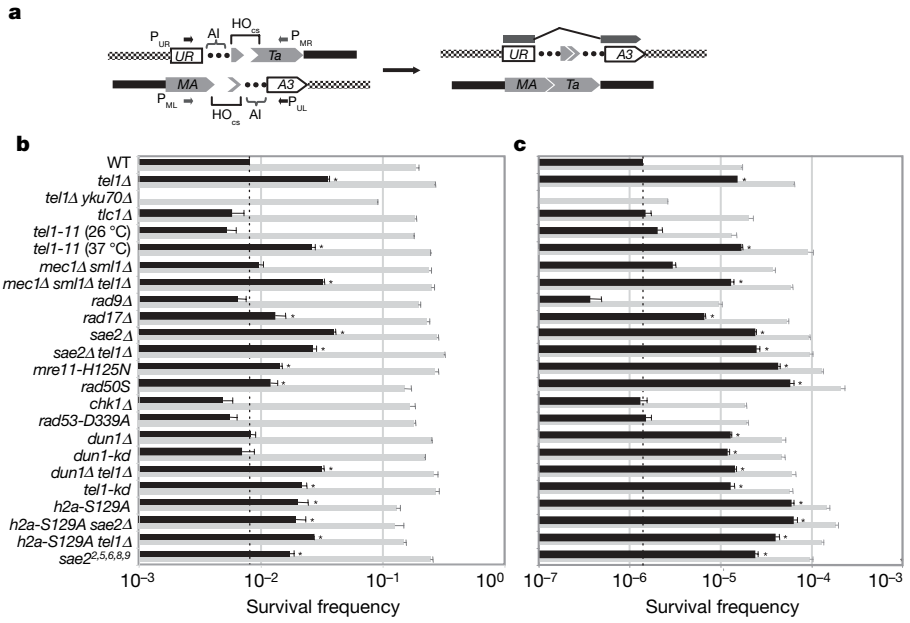


Figure 1 | The Tel1 pathway suppresses chromosome translocations by NHEJ. a, Diagram of the construct in *S. cerevisiae* carrying two HO recognition sites (HO_{cs}), a part of the *URA3* gene and the intronic sequence (AI). **b**, **c**, Yeast strains carrying the construct in **a** experienced two simultaneous DSBs by expression of HO for 1 h in liquid culture (**b**) or by plating onto galactose-containing media (**c**). Frequency of all survivors (GAL^R, grey bars) and uracil prototrophs among them (URA^+, black bars; representing chromosome translocations) are plotted on a logarithmic scale. Data represent the mean ± s.d. from three or more independent experiments. Values significantly higher ($P < 0.001$ by student's *t*-test) than Ura+ survival frequency from the wild-type (WT) strain are marked with an asterisk (see Supplementary Table 2 for details). Locations of primers used to measure kinetics of DSB induction and end joining in Supplementary Figs 1 and 5 are shown.

Because the *mre11* mutation that abrogates its nuclease activity enhances chromosomal translocations after persistent HO induction (Fig. 1c), reduced end processing may increase the overall NHEJ efficiency and the number of NHEJ-mediated chromosome translocations[22]. Indeed, deletion of *SAE2*, *TEL1* or expression of *tel1-kd, hta1-S129A hta2-S129A*, or *sae2*^{2,5,6,8,9} all reduced formation of ssDNA and RPA binding at 0.2-, 1.0-, 1.6- or 5.0-kb from a DSB, as detected by quantitative amplification of ssDNA (QAOS) (Fig. 2b, d) or chromatin immunoprecipitation (Fig. 2a, c and Supplementary Fig. 6)[23]. The results suggest that Tel1- and Mre11-dependent end processing suppresses NHEJ and chromosome translocations.

Deletion of *SAE2* or *TEL1* also stimulates a three- to fourfold increase in chromosome translocation among survivors (Supplementary Table 2). Additional mechanism(s) must exist to suppress end joining between different chromosomes. In yeast, broken chromosome ends remain associated with each other at G2 for up to 6–8 h after DSB induction in a Rad50-, Sae2- and Rad52-dependent manner[24,25]. We propose that intrachromosomal association/tethering of DNA ends may suppress chromosome translocations. To test this premise, donorless yeast strains with LacO arrays located at 50 kb on both sides of the HO break on chromosome VII and expressing LacI–GFP fusion proteins that bind LacO (Fig. 3a)[24] were arrested at G1 and assessed for end tethering when NHEJ is most active[22].

Approximately 16% of *TEL1*+ cells induced to a DSB by HO endonuclease contained GFP foci separated by greater than 0.5 μm, suggesting that the ends of a broken chromosome are tethered to maintain intrachromosomal association (Fig. 3b, d). In contrast, deletion of *TEL1* or *MRE11* increased the separation of two broken chromosome ends by ~3-fold (Fig. 3c, d), whereas Rad52 is dispensable for end tethering at G1. Expression of *tel1-kd* or *sae2*^{2,5,6,8,9} also promoted dissociation of DNA ends after a DSB induction (Fig. 3c and Supplementary Fig. 7). The results are consistent with Tel1 and Sae2 suppressing NHEJ-mediated reciprocal chromosome translocations by facilitating intrachromosomal association of broken DNA ends.

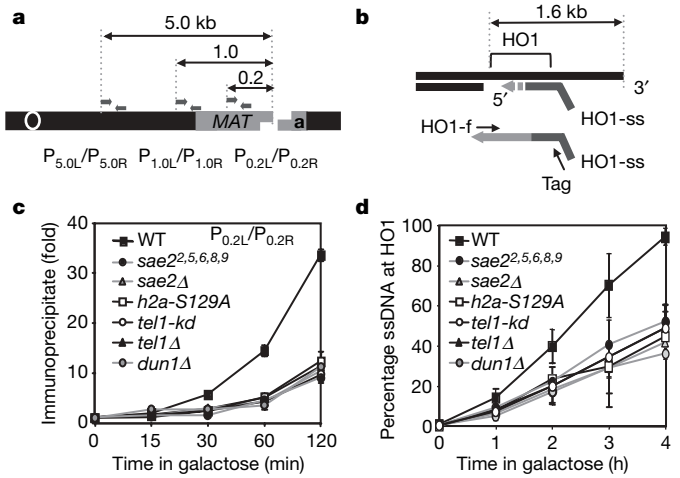


Figure 2 | Tel1- and Sae2-mediated end processing suppresses mutagenic NHEJ and chromosome translocations. ChIP assays were used to assess the levels of RPA at the DSB in *dun1*Δ, *h2a-S129A*, *sae2*Δ, *sae2*^{2,5,6,8,9}, *tel1*Δ and *tel1-kd* strains using an anti-RPA antibody as described in Methods. **a**, The positions of the HO cut site and the location of primers. **b**, QAOS assay that measures the amount of ssDNA at 1.6-kb centromere proximal (HO1) to a DSB site by primer extension at non-denaturing conditions using tagging primer HO1-ss, followed by real-time quantitative PCR using Tag and HO1-f primers[23]. **c**, PCR signals from a primer set that anneals 0.2-kb distal to the HO break at different durations of HO expression were quantified and plotted as a graph. Fold immunoprecipitate represents the ratio of the RPA PCR signal before and after HO induction, normalized by the PCR signal of the *PRE1* control. **d**, Percentage ssDNA was calculated by dividing the PCR signals using primer HO1-f and Tag with those from a denatured sample at each time point and plotted as a graph. Each point is the average ± s.d. from two separate experiments.

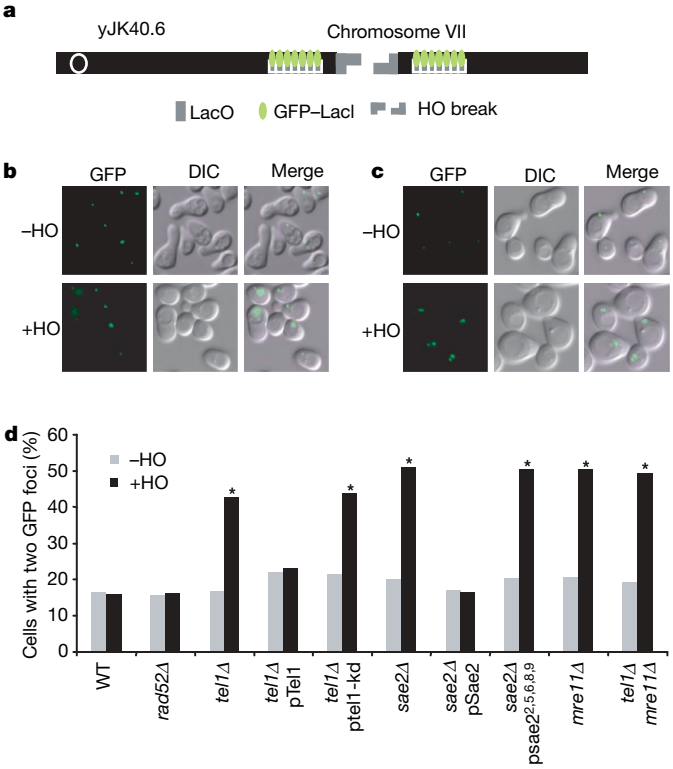


Figure 3 | Tel1 and Sae2 contribute to intrachromosomal association. **a**, Yeast cells expressing LacI–GFP which binds two LacO arrays located ~50 kb on either side of the HO break were induced for HO expression at G1. **b**, **c**, Samples taken at 3 h after HO induction were analysed for end tethering. Separation of two GFP–LacI foci by more than 0.5 μm is considered to indicate lack of end tethering. Typically, the separation distance between foci was measured for 300–1,000 cells per sample and the results are plotted in **d**. Results from *tel1*Δ, *sae2*Δ, but expressing Tel1 (pTel1), tel1-kd (ptel1-kd), Sae2 (pSae2) or sae2^{2,5,6,8,9} (psae2^{2,5,6,8,9}) from a centromeric plasmid as well as *mre11*Δ, or *rad52*Δ, are shown. Values significantly higher ($P < 0.01$ by student's *t*-test on pairwise comparisons of the percentage of cells with two GFP foci before and after galactose addition) are marked with an asterisk

We have demonstrated that a cell possesses ways in which to discern which ends to join and thus suppress end joining between different chromosomes. In mammals, deficiencies in ATM, H2AX, or the Mre11 complex are all associated with increased chromosome abnormalities and predisposition to cancers[26–28]. Our findings reveal the mechanism that eukaryotic cells use to avoid chromosome translocations, and provide a possible explanation as to why ATM-deficient cells accumulate chromosome translocations.

METHODS SUMMARY

Pulse-field gel electrophoresis[29], chromatin immunoprecipitation[30] and intrachromosomal association assays[24] were performed as described previously.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 17 January; accepted 30 April 2008.

1. Zhang, Y. & Rowley, J. D. Chromatin structural elements and chromosomal translocations in leukemia. *DNA Repair* **5**, 1282–1297 (2006).
2. Lieber, M. R., Yu, K. & Raghavan, S. C. Roles of nonhomologous DNA end joining, V(D)J recombination, and class switch recombination in chromosomal translocations. *DNA Repair* **5**, 1234–1245 (2006).
3. Putnam, C. D., Pennaneach, V. & Kolodner, R. D. *Saccharomyces cerevisiae* as a model system to define the chromosomal instability phenotype. *Mol. Cell. Biol.* **25**, 7226–7238 (2005).

4. Myung, K., Datta, A. & Kolodner, R. D. Suppression of spontaneous chromosomal rearrangements by S phase checkpoint functions in *Saccharomyces cerevisiae*. *Cell* **104,** 397–408 (2001).
5. Haber, J. E. Transpositions and translocations induced by site-specific double-strand breaks in budding yeast. *DNA Repair* **5,** 998–1009 (2006).
6. Haber, J. E. & Leung, W. Y. Lack of chromosome territoriality in yeast: promiscuous rejoining of broken chromosome ends. *Proc. Natl Acad. Sci. USA* **93,** 13949–13954 (1996).
7. Khanna, K. K. Cancer risk and the ATM gene: a continuing debate. *J. Natl Cancer Inst.* **92,** 795–802 (2000).
8. Greenwell, P. W. *et al.* TEL1, a gene involved in controlling telomere length in *S. cerevisiae*, is homologous to the human ataxia telangiectasia gene. *Cell* **82,** 823–829 (1995).
9. Usui, T., Ogawa, H. & Petrini, J. H. A. DNA damage response pathway controlled by Tel1 and the Mre11 complex. *Mol. Cell* **7,** 1255–1266 (2001).
10. Chakhparonian, M., Faucher, D. & Wellinger, R. J. A mutation in yeast Tel1p that causes differential effects on the DNA damage checkpoint and telomere maintenance. *Curr. Genet.* **48,** 310–322 (2005).
11. Ritchie, K. B., Mallory, J. C. & Petes, T. D. Interactions of TLC1 (which encodes the RNA subunit of telomerase), TEL1, and MEC1 in regulating telomere length in the yeast *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **19,** 6065–6075 (1999).
12. Nyberg, K. A., Michelson, R. J., Putnam, C. W. & Weinert, T. A. Toward maintaining the genome: DNA damage and replication checkpoints. *Annu. Rev. Genet.* **36,** 617–656 (2002).
13. Zhao, X., Muller, E. G. & Rothstein, R. A suppressor of two essential checkpoint genes identifies a novel protein that negatively affects dNTP pools. *Mol. Cell* **2,** 329–340 (1998).
14. Clerici, M., Mantiero, D., Lucchini, G. & Longhese, M. P. The *Saccharomyces cerevisiae* Sae2 protein promotes resection and bridging of double strand break ends. *J. Biol. Chem.* **280,** 38631–38638 (2005).
15. Kim, H. S. *et al.* Functional interactions between Sae2 and the Mre11 complex. *Genetics* **178,** 711–723 (2008).
16. Moreau, S., Ferguson, J. R. & Symington, L. S. The nuclease activity of Mre11 is required for meiosis but not for mating type switching, end joining, or telomere maintenance. *Mol. Cell. Biol.* **19,** 556–566 (1999).
17. Alani, E., Padmore, R. & Kleckner, N. Analysis of wild-type and rad50 mutants of yeast suggests an intimate relationship between meiotic chromosome synapsis and recombination. *Cell* **61,** 419–436 (1990).
18. Bashkirov, V. I., Bashkirova, E. V., Haghnazari, E. & Heyer, W. D. Direct kinase-to-kinase signaling mediated by the FHA phosphoprotein recognition domain of the Dun1 DNA damage checkpoint kinase. *Mol. Cell. Biol.* **23,** 1441–1452 (2003).
19. Mallory, J. C. *et al.* Amino acid changes in Xrs2p, Dun1p, and Rfa2p that remove the preferred targets of the ATM family of protein kinases do not affect DNA repair or telomere length in *Saccharomyces cerevisiae*. *DNA Repair* **2,** 1041–1064 (2003).
20. Shroff, R. *et al.* Distribution and dynamics of chromatin modification induced by a defined DNA double-strand break. *Curr. Biol.* **14,** 1703–1711 (2004).
21. Baroni, E., Viscardi, V., Cartagena-Lirola, H., Lucchini, G. & Longhese, M. P. The functions of budding yeast Sae2 in the DNA damage response require Mec1- and Tel1-dependent phosphorylation. *Mol. Cell. Biol.* **24,** 4151–4165 (2004).
22. Ira, G. *et al.* DNA end resection, homologous recombination and DNA damage checkpoint activation require CDK1. *Nature* **431,** 1011–1017 (2004).
23. van Attikum, H., Fritsch, O. & Gasser, S. M. Distinct roles for SWR1 and INO80 chromatin remodeling complexes at chromosomal double-strand breaks. *EMBO J.* **26,** 4113–4125 (2007).
24. Kaye, J. A. *et al.* DNA breaks promote genomic instability by impeding proper chromosome segregation. *Curr. Biol.* **14,** 2096–2106 (2004).
25. Lobachev, K., Vitriol, E., Stemple, J., Resnick, M. A. & Bloom, K. Chromosome fragmentation after induction of a double-strand break is an active process prevented by the RMX repair complex. *Curr. Biol.* **14,** 2107–2112 (2004).
26. Bassing, C. H. & Alt, F. W. H2AX may function as an anchor to hold broken chromosomal DNA ends in close proximity. *Cell Cycle* **3,** 149–153 (2004).
27. Celeste, A. *et al.* Genomic instability in mice lacking histone H2AX. *Science* **296,** 922–927 (2002).
28. Petrini, J. H. The Mre11 complex and ATM: collaborating to navigate S phase. *Curr. Opin. Cell Biol.* **12,** 293–296 (2000).
29. Unal, E. *et al.* DNA damage response pathway uses histone modification to assemble a double-strand break-specific cohesin domain. *Mol. Cell* **16,** 991–1002 (2004).
30. Shim, E. Y. *et al.* RSC mobilizes nucleosomes to improve accessibility of repair machinery to the damaged chromatin. *Mol. Cell. Biol.* **27,** 1602–1613 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to S. Brill, J. Haber, W. D. Heyer, M. P. Longhese, J. L. Ma, T. Paull, E-Y. Shim, P. Sung, D. Toczyski, R. Wellinger and members of the S. Lee laboratory for reagents and helpful suggestions. This work was supported by grants in the National Institutes of Health and the Texas Advanced Research program to S.E.L. S.E.L. is a Leukemia and Lymphoma Society Scholar.

Author Contributions K.L., Y.Z. and S.E.L. designed experiments. K.L. performed the chromosome translocation assay shown in Fig. 1, Supplementary Table 2 and Supplementary Figs 1–6. K.L. carried out chromatin immunoprecipitation and QAOS assays in Fig. 2. Y.Z. analysed intrachromosomal end association in Fig. 3, along with Supplementary Fig. 7. S.E.L. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to S.E.L. (lees4@uthscsa.edu).

METHODS

Strains. All strains are derivatives of JKM179, which has the genotype *hoΔ MATα hmlΔ::ADE1 hmrΔ::ADE1 ade1-100 leu2-3,112 lys5 trp1::hisG' ura3-52 ade3::GAL::HO* (Supplementary Table 1). The SLY60 strain was generated by replacing *ura3-52* with a 1.2-kb *URA3* fragment interrupted by a 117-bp HO cleavage site and an artificial intron containing a splicing consensus sequence at the EcoRV site using one-step gene replacement, and recovered from FOA-resistant, uracil auxotrophs after an hour HO induction and plating onto minimal media containing 5′-FOA. Yeast strains yJK40.6, 37.2 and 98.9 used in the intrachromosomal association assay were a gift from D. Toczyski[24].

Chromatin immunoprecipitation. Chromatin immunoprecipitation assays were performed as described previously[30]. After immunoprecipitation and reverse cross-linking, purified DNA was analysed by real-time quantitative PCR using three sets of primers that anneal 0.2 kb, 1.0 kb and 5.0 kb from the DSB, as well as primers specific for the *PRE1* gene situated on chromosome V as a control. The antibodies for RPA were a gift from S. Brill.

Intrachromosomal association assay. The yeast strain (yJK40.6) harbouring multiple repeats of LacO on both sides of an HO break and expressing a GFP–LacI fusion protein was grown in YEP-glycerol at 30 °C to a logarithmic culture (10^7 cells ml^{-1}) and arrested in G1 by addition of 15 μg ml^{-1} α-mating factor to the culture medium[24]. HO breaks were induced by the addition of galactose to a final concentration of 2% (v/v) for 3 h. GFP–LacI was induced for 1 h by supplementing media with $CuSO_4$ to 50 μM and cells were photographed using a Zeiss Axioplan II imaging system. The two arrays of LacO are separated by a distance of only 50 kb so that only one GFP focus is observed if chromosome VII is intact or the two broken ends are tethered together after an HO break. Two GFP foci are detected when the two broken ends fall apart. Serial images were captured in 0.25-μm increments over 2 μm in the z plane. The z-series images were deconvolved using the AV4 Mod deconvolution software (Zeiss).

Pulsed-field gel electrophoresis. Approximately 5×10^7 logarithmic cells cultured in YPD medium were harvested and fixed with 70% ethanol at 4 °C overnight. Agarose-embedded yeast DNA plugs were prepared using a CHEF genomic DNA plug kit according to the manufacturer's protocol and applied to a CHEF DR III system (BioRad).

naturejobs

THE CAREERS MAGAZINE FOR SCIENTISTS

At a recent science meeting in San Diego, some young scientists asked me to identify significant job trends. I replied that industrial employers, especially in the pharmaceutical sector, are increasingly seeking employees with master's degrees tailored to their specific business needs. "So is this the death of the PhD?" one of them asked provocatively, noting the paucity of university positions. Well no, clearly not.

Nevertheless, this rising demand for master's degrees could act as a springboard for a qualification in the United States called the professional science master's (PSM) degree (see *Nature* **445,** 458; 2007). This offers a blend of team-building and communication skills, legal and regulatory information and business savvy that employers tend to like. And as it takes only two to three years to complete, it is attractive to those scientists wary of spending years on a PhD and perhaps a postdoc, only to fail to reach the desired career outcome.

On 11 July, the PSM was endorsed by a US National Academies panel, which saw a niche for it in the pharmaceutical, biotechnology and defence sectors. The panel backed the idea of degrees that foster communication and business skills in addition to science know-how. Although conventional master's in computer sciences and geosciences often offer a clear path to the workplace, the broad curricula in physics, biology and chemistry often provide fewer practical skills for jobs outside academia.

The PSM is not an inevitable runaway hit. The degree programmes must produce graduates who have success in industry and thereby make the PSM a reputable choice for future students. And there is the issue of cost. In the United States, PhD training is often paid for with grant money or fellowships. As a result, some life-sciences graduates earn their conventional master's for free after declining to continue with their PhD. Many students will have to pay for the PSM — but is it worth it? If it quickly leads to a well paid job, it certainly might be. Scientists and employers seem to agree that a new class of training is in order. Now students and institutions have to find a way to ensure that it's worth the investment.

Gene Russo is editor of *Naturejobs*.

CONTACTS
Editor: Gene Russo

European Head Office, London
The Macmillan Building,
4 Crinan Street, London N1 9XW, UK
Tel: +44 (0) 20 7843 4961
Fax: +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

European Sales Manager:
Andy Douglas (4975)
e-mail: a.douglas@nature.com
Business Development Manager:
Amelie Pequignot (4974)
e-mail: a.pequignot@nature.com
***Nature*events:**
Claudia Paulsen Young (+44 (0) 20 7014 4015)
e-mail: c.paulsenyoung@nature.com
France/Switzerland/Belgium:
Muriel Lestringuez (4994)
Southwest UK/RoW: Nils Moeller (4953)

Scandinavia/Spain/Portugal/Italy:
Evelina Rubio-Hakansson (4973)
Northeast UK/Ireland:
Matthew Ward (+44 (0) 20 7014 4059)
North Germany/The Netherlands:
Reya Silao (4970)
South Germany/Austria:
Hildi Rowland (+44 (0) 20 7014 4084)

Advertising Production Manager:
Stephen Russell
To send materials use London address above.
Tel: +44 (0) 20 7843 4816
Fax: +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com
***Naturejobs* web development:** Tom Hancock
***Naturejobs* online production:** Dennis Chu

US Head Office, New York
75 Varick Street, 9th Floor,
New York, NY 10013-1917
Tel: +1 800 989 7718
Fax: +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

India
Vikas Chawla (+91 1242881057)
e-mail: v.chawla@nature.com

Japan Head Office, Tokyo
Chiyoda Building, 2-37 Ichigayatamachi,
Shinjuku-ku, Tokyo 162-0843
Tel: +81 3 3267 8751
Fax: +81 3 3267 8746

Asia-Pacific Sales Manager:
Ayako Watanabe (+81 3 3267 8765)
e-mail: a.watanabe@natureasia.com
Business Development Manager, Greater China/Singapore:
Gloria To (+852 2811 7191)
e-mail: g.to@natureasia.com

NATURE|Vol 454|24 July 2008

MOVERS

Brent Reynolds, director of the Adult Stem Cell Engineering and Therapeutic Core, McKnight Brain Center, University of Florida, Gainesville

2004–08: Visiting scientist, then senior research fellow, Queensland Brain Institute, University of Queensland, Brisbane, Australia
1992–98: Director, then vice-president and director of research, NeuroSpheres, Calgary, Canada

The price of having too much too soon can be high. When Brent Reynolds isolated and cultured mouse brain cells during his PhD at the University of Calgary in Canada, his career was on the fast track. He eschewed a postdoc to start his own company with adviser Sam Weiss, a neuroscience professor at the university. But after Ciba-Geigy invested in the company, a merger and subsequent divestiture effectively sidelined Reynolds and Weiss's technology. Reynolds went from having his research featured on the cover of a major scientific journal to wondering if he wanted to stay in science at all.

He decided to take a career break. It lasted six years. "I needed some perspective," Reynolds says. "I had become somewhat disillusioned with science, and in particular with science and business." Returning to an earlier interest in Eastern medicine and philosophy, Reynolds devoted himself to studying Chinese medicine, acupuncture and yoga. He established a yoga studio in Thailand, then moved to Salt Spring Island off the coast of Vancouver, where there were "no bridges, no highways, no traffic lights, no parking garages and, importantly, no parking meters".

But his move away actually brought Reynolds back into the scientific fold. At a yoga training course, he met an instructor who called Australia "the coolest place in the world". A few weeks later, former colleague Rod Rietze invited Reynolds to join him at the new Queensland Brain Institute in Brisbane, and his science career took off again.

After a few years in Australia, Reynolds decided to return to North America and join those attempting to translate stem-cell technology into clinical treatments. Dennis Steindler, director of the McKnight Brain Institute in Florida, had admired the neural stem-cell work Reynolds conducted as a graduate student. "His insights are incredible in terms of how to grow them and how to get them to behave," Steindler says. He believes Reynolds's work will help McKnight researchers use stem cells in new brain-cancer therapies. Reynolds expects that his familiarity with the holistic approach of Eastern traditions may help improve the often too-reductionist approach to stem-cell research.

Eastern philosophy also validated his decisions to take a break from science and to return to it. The Taoist concept of *wu wei*, practically applied, advocates following instincts or hunches, Reynolds notes. Steindler hopes these instincts will lead the institute closer to stem-cell therapies. ■

Paul Smaglik

NETWORKS & SUPPORT

Vitae for postgraduate development

Postgraduate scientists in Britain have a new national programme devoted to personal and career development. Vitae — the latest incarnation of the UK GRAD Programme — will cater to both postgraduate students and postdocs (often called 'research staff' in Britain), whose needs, notably job security, often get overlooked by government and institutions.

With 4,500 members, Vitae wants to become academia's national policy instrument for research career development, and to get employers and postgrads talking constructively.

It recently released an updated concordat, with seven principles spelling out expectations and responsibilities. These include employers recognizing the need to retain and value good researchers, and researchers sharing the responsibility and taking up lifelong-learning opportunities. Endorsed by funders, universities, professional societies and the European Commission, the concordat is a clear statement that the development of researchers is as important as research output, says Vitae chair Janet Metcalfe. She notes that many currently end up with an unsettling series of two-year contracts.

Researchers' independence deserves attention as well, says John Bothwell, co-founder of the UK National Research Staff Association. Instead of spending five to ten years working on a mentor's ideas, Bothwell suggests that young investigators need targeted funding to develop their own research. Vitae may help such proposals by establishing metrics for fledgling scientists' accomplishments that could be used by funding bodies.

"The most effective thing Vitae could do is encourage postdocs to look up from the lab bench and think about their career direction beyond simply publishing papers," says Bothwell. Vitae's website (www.vitae.ac.uk) has links to interactive tools to assess sector-specific skill competencies as well as discussion forums. Its 2008 programme of events includes three- and four-day courses to help postgraduates learn to manage career choices, and two-day Effective Researcher workshops on project management and leadership skills. Vitae will also host networking days for researchers interested in the drug or biotechnology industries.

"Vitae is one of several ways to ensure that Britain's research infrastructure stays world class, even in the face of competition from places such as China and India," Metcalfe says. ■

Virginia Gewin

POSTDOC JOURNAL

An unwelcome intrusion

Western civilization is staring me in the face again, and I'm not sure I like it. By the time this journal entry is published, I'll be hailing taxis in the muddy streets of Addis Ababa and hassling officials for the visa that will allow me back into the United States. I'll be out of my little comfort zone.

For two months I'll be away from the Simien Mountains, analysing data and 'catching my breath' back in the United States. Although I've been missing Western food, company, daily showers and little luxuries like tarred roads, I suddenly don't feel ready to return to Western culture.

My stomach clenches when I think that this is my last week in the mountains. I have a lot of work to do — back-ups of data, last-minute experiments, organizing things so that our permanent field assistant can continue his work until my return in September.

But I'm not feeling that stressed about the work that remains. What I dread is reverse culture shock.

I'm trying to get my mind around the big change about to happen. I'll have a car, watch movies, see lots of friends, eat to my heart's content ... These are good things. And I'll be surrounded by tall buildings instead of massive cliff walls, fashion-conscious students instead of crooked-toothed scouts, city lights instead of stars. I think I need a hug. ■

Aliza le Roux is a postdoctoral fellow in animal behaviour at the University of Michigan.

Hillcrest v. *Velikovsky*

An act of God?

Peter Watts

The facts of the case were straightforward. Lacey Hillcrest of Pensacola, 50 years old and a devout Pentecostal, had been diagnosed with inoperable lymphatic cancer and given six months to live. Five years later she was still alive, albeit frail. She attributed her survival to a decorative silver-plated cross received from her sister, Gracey Balfour. Witnesses attested that Mrs Hillcrest's condition improved dramatically upon acquisition of the totem, a product of the Graceland Mint alleged to contain an embedded fragment of the original Crucifix of Golgotha.

On the morning of 27 June, Mrs Hillcrest and her sister patronized the Museum of Quackery and Pseudoscience, owned and managed by one Linus C. Velikovsky. The museum contained a variety of displays concerning discredited beliefs, theories and outright hoaxes perpetrated throughout American history. Mrs Balfour entered into a heated discussion with another museum patron at the Intelligent Design exhibit, temporarily losing track of her sister; they eventually reconnected at a display concerning psychosomatic phenomena, specifically placebo effects and faith healing. Mrs Hillcrest had evidently spent some time perusing the display and was subsequently described as 'subdued and uncommunicative'. Within a month she was dead.

The charge against Mr Velikovsky was negligent homicide.

The Prosecution called Dr Andrew deTritus, a clinical psychologist with an impressive record of expert testimony on any (and sometimes conflicting) sides of a given issue. Dr deTritus testified to the uncontested reality of the placebo effect, pointing out that 'attitude' and 'outlook' — like any other epiphenomenon of the brain — were ultimately neurochemical in nature. *Belief* literally rewired the brain, and the existence of placebo effects showed that such changes could have a real impact on human health.

Velikovsky took the stand in his own defence, which was straightforward: all claims presented by his displays were factually accurate and supported by scientific evidence. The prosecution objected to this point on the grounds of relevance but was, after some discussion, overruled.

Far from disputing Velikovsky's claims during cross-examination, however, the Prosecution used them to bolster its own case. The defendant had deliberately set up shop in "one of our great country's most devout regions, with no thought to the welfare of the Lacey Hillcrests of the world". By his own admission, Mr Velikovsky had chosen Florida "because of all the creation museums", and had clearly been intent on rubbing people's noses in the alleged falsity of their beliefs. Furthermore, Mr Velikovsky was obviously well-versed in placebo effects, having built an erudite display on the subject. What did he *think* would happen, the Prosecution thundered, when he forced his so-called *truth* down the throat of someone whose motto — knitted into her favourite throw-cushion — was *If ye have faith the size of a mustard seed, ye shall move mountains*? In telling 'the truth' Velikovsky had knowingly and recklessly endangered the very *life* of another human being.

JACEY

Velikovsky pointed out that he hadn't even known Lacey Hillcrest existed, adding that needlepointing something onto a pillowcase did not necessarily make it true. The Prosecution responded that the man who plants landmines in a playground doesn't know the names of his victims either, and asked if the defendant's needlepoint remark meant that he was now calling Jesus a liar. The Defence objected repeatedly throughout.

The Defence had, in fact, fought an uphill battle ever since her client's swearing-in, during which Velikovsky had asked whether swearing to tell the truth on "a book of falsehoods" might undermine the court's alleged devotion to empiricism. The jury had seemed unimpressed by that question, and did not seem to have subsequently become more sympathetic.

Perhaps, if worst came to worst, their verdict might be set aside on technical grounds. But the closest thing to a precedent the Defence could unearth was *Dexter* v. *HerpBGone*, involving a mail-order scheme in which a mixture of sugar and baking soda had been marketed as a cure for herpes at $200/treatment. Although this 'cure' had (unsurprisingly) proven ineffective, HerpBGone's council had cited Waber *et al.* 2008 (ref. 1) — which clearly showed that a placebo's efficacy increased with price — arguing that the treatment *could* have worked if Dexter had only paid more for it. As he had refused to do so (the same product was sold under a different name at $4,000), responsibility devolved to the plaintiff. The case had been dismissed.

It would have been a risky gambit. The parallels were far from exact. Instead, the Defence recalled Grace Balfour to the stand and asked whether she believed the Bible to be the revealed Word of God. Mrs Balfour readily conceded as much. It was her faith, she maintained, that allowed her to stay strong when that horrible man at the Creation display had mocked her with his talk of monkeymen and radioisotopes. She had seen fossils for what they *truly* were, the tests of faith described in Deuteronomy 13.

Asked then why her sister evidently did not share her strength of belief, Mrs Balfour allowed— somewhat reluctantly— that "that horrid little Russian" had shattered her sister's faith with his "lies and deceit".

But did not the Bible itself arm the faithful against such wickedness? Did not Matthew warn that "false prophets shall rise, and deceive many"? Could Second Peter have *been* any more explicit than "There shall be false teachers among you, who shall bring in damnable heresies"?

Well, yes, Mrs Balfour allowed. Certainly, Velikovsky was a False Prophet. Sadly, as the Defence reminded her, false prophecy was not a criminal offence.

Ultimately there was no need to resort to technical exemptions. The jury, having been presented with the facts of the case, was unanimous: Lacey Hillcrest had not shown the courage for their conviction. Whose fault was it, after all, that her faith had been so much smaller than a mustard seed? ■

1. Waber, R. L., Shiv, B., Carmon, Z. & Ariely. D. *J. Am. Med. Assoc.* **299,** 1016–1017 (2008).

Peter Watts's first story for Futures failed to provoke the desired howls of envious outrage from former colleagues who'd sneered at his decision to leave academia and write science fiction, and who then spent years trying desperately to get published in *Nature*. He hopes that more satisfying outbursts will result from repeated publication.